



DAGSTUHL REPORTS

Volume 15, Issue 10, October 2025

Trustworthy Evidence-Based Elections (Dagstuhl Seminar 25411) <i>Josh Benaloh, Peter Rønne, and Melanie Volkamer</i>	1
Creativity, GenAI, and Software Development: A Future Together (Dagstuhl Perspectives Workshop 25412) <i>Victoria Jackson, Rafael Prikladnicki, Daniel Russo, Margaret-Anne Storey, and André van der Hoek</i>	23
Sound Static Program Analysis in Modern Software Engineering (Dagstuhl Seminar 25421) <i>Pietro Ferrara, Liana Hadarean, Jorge A. Navas, Caterina Urban, and Greta Dolcetti</i>	37
Cognitive Sensing and Interaction (Dagstuhl Seminar 25422) <i>Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner</i>	75
Quantum Cryptanalysis (Dagstuhl Seminar 25431) <i>Gorjan Alagic, Simona Etinski, Stacey Jeffery, and Rainer Steinwandt</i>	105
Deep Continual Learning in the Foundation Model Era (Dagstuhl Seminar 25432) <i>Christopher Kanan, Martin Mundt, Tinne Tuytelaars, Joost van de Weijer, and Timm Felix Hess</i>	119
Competitions and Empirical Evaluations in Automated Reasoning (Dagstuhl Seminar 25441) <i>Johannes K. Fichte, Matti Järvisalo, Aina Niemetz, Andreas Niskanen, and Guido Tack</i>	135
Augmenting Human Creativity with AI (Dagstuhl Seminar 25442) <i>Tilman Dingler, Elisa Giaccardi, Steven Rick, Irina Shklovski, and Samangi Wadinambiarachchi</i>	155

ISSN 2192-5283

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at <https://www.dagstuhl.de/dagpub/2192-5283>

Publication date

June, 2026

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <https://dnb.d-nb.de>.

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC BY 4.0).



In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:

- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Aims and Scope

The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.

In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:

- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Editorial Board

- Elisabeth André
- Franz Baader
- Goetz Graefe
- Reiner Hähnle
- Barbara Hammer
- Lynda Hardman
- Steve Kremer
- Rupak Majumdar
- Heiko Mantel
- Lennart Martens
- Albrecht Schmidt
- Wolfgang Schröder-Preikschat
- Holger Hermanns (*Editor-in-Chief*)
- Heike Wehrheim
- Verena Wolf
- Martina Zitterbart

Editorial Office

Michael Wagner (*Managing Editor*)
Michael Didas (*Managing Editor*)
Jutka Gasiorowski (*Editorial Assistance*)
Dagmar Glaser (*Editorial Assistance*)
Thomas Schillo (*Technical Assistance*)
Christina Schwarz (*Editorial Assistance*)

Contact

Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de

<https://www.dagstuhl.de/dagrep>

Digital Object Identifier: 10.4230/DagRep.15.10.i

Trustworthy Evidence-Based Elections

Josh Benaloh^{*1}, Peter Rønne^{*2}, and Melanie Volkamer^{*3}

1 Microsoft Research – Redmond, US. benaloh@microsoft.com

2 University of Luxembourg, LU. peter.roenne@uni.lu

3 KIT – Karlsruher Institut für Technologie, DE. melanie.volkamer@kit.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25411 “Trustworthy Evidence-Based Elections”. The seminar brought together 41 participants mainly from computer science, but also sociology and statistics, including members of organizations, industry, and national agencies; all experts on elections and secure voting. The five-day seminar had a full program primarily with talks presenting the latest research results, challenges, and observations including ample time for discussions. However, the presentation sessions were complemented with demo sessions and daily breakout groups continuing the most interesting discussions that emerged from the talks.

Seminar October 5–10, 2025 – <https://www.dagstuhl.de/25411>

2012 ACM Subject Classification Security and privacy → Cryptography; Security and privacy → Human and societal aspects of security and privacy; Applied computing → Voting / election technologies

Keywords and phrases elections, risk-limiting audits, Trust, usable security, verifiable voting, voting

Digital Object Identifier 10.4230/DagRep.15.10.1

1 Executive Summary

Josh Benaloh (Microsoft Research – Redmond, US)

Peter Rønne (University of Luxembourg, LU)

Melanie Volkamer (KIT – Karlsruher Institut für Technologie, DE)

License  Creative Commons BY 4.0 International license
© Josh Benaloh, Peter Rønne, and Melanie Volkamer

This 5-day Dagstuhl Seminar 25411 was dedicated to “Trustworthy Evidence-Based Elections”, a topic whose relevance and importance is witnessed by a number of democracies being under threat and elections becoming political battlegrounds in themselves rather than free and fair elections of political candidates. In this earnest backdrop, the seminar reviewed the state of art of secure elections, explored paths towards trustworthy elections, and the open challenges, in particular by sharing and elaborating on novel ideas, approaches, and use cases.

This report gives a description of the program, collects the abstracts of the talks in Sec. 3, and a selection of the breakout groups in Sec. 4.

Day 1

The Dagstuhl Seminar was opened with a round of short individual presentations, followed by a discussion about the organization of the seminar and the planning of topics and talks. All participants were allowed and encouraged to contribute with talks on recent research, election

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license
Trustworthy Evidence-Based Elections, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 1–22
Editors: Josh Benaloh, Peter Rønne, and Melanie Volkamer



Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

experiences, open problems, and ideas. In addition, it was agreed to end each day with breakout sessions and a final joint synthesis session with reports from each of the breakout groups. Due to the large number of individual talks, the breakout groups had to be limited to one-two sessions per day. Each talk was followed by active discussions.

The first day also managed to feature two sessions of talks, first a short session on Evidence-Based Elections and a longer one on Usability in voting. The breakout groups included a group on Commitment-Based Elections, Coercion-Resistance & Usability, and Education & Outreach.

Day 2

The second day had two sessions of talks on Cryptographic Voting Protocols and Risk-Limiting Audits. Furthermore, in the afternoon a demo session was held which presented an explanatory video¹ to enable informed decisions on whether or not to go for Internet voting, a demo of the Selene voting scheme, a demo on a protocol for voter enrollment in coercion-resistant elections, and a system to test paper ballots before elections in particular for logical errors. Finally, a demo on visualising the counting of individual ballots in STV was postponed until day 4.

The breakout groups included groups on Universally Verified elections and RLAs.

Day 3

The morning of day three had two sessions of talks on respectively Trust and Security Definitions. Further, it included a talk on e-collection of signatures for referendums, a topic sharing many security problems with e-voting.

In the afternoon, the participants split into two groups, one group explored the hiking trails around Schloss Dagstuhl while another group visited the Celtic Hillfort of Otzenhausen.

Day 4

Day four had three sessions: Cryptographic Voting Protocols (the second session on this topic), Verified Implementations & Formal Verification, and Trust & Elections. From the talks in these sessions, several breakout groups emerged, especially groups looking at boardroom voting, trust in voting, a group initiating the work towards the creation of an open source software developer kit (SDK) for e-voting, and trustworthy randomness-generation for election audits.

Day 5

The final day started with discussions on creating a white-paper on trustworthy evidence-based elections, the structure of this and responsible editors. In addition, two rounds of breakout groups were organized: both dedicated to new topics, following up on earlier groups and preparing the whitepaper topics. In particular, there were groups on Usability, Post-Quantum E-Voting, Formal Verification and Security Definitions.

¹ https://secuso.org/2025-07-E-Voting-Explanation-Video/Video/paper_version/final_long_subtitles.mp4

2 Table of Contents

Executive Summary

<i>Josh Benaloh, Peter Rønne, and Melanie Volkamer</i>	1
--	---

Overview of Talks

My Journey to Commitment	
<i>Josh Benaloh</i>	6
Risk-limiting Audits for Complex Elections: Progress and What's Next	
<i>Michelle Blom, Alexander Ek, and Vanessa Teague</i>	6
Breaking and fixing CHVote, a protocol for the Swiss context	
<i>Véronique Cortier</i>	6
A Practical and Fully Distributed E-Voting Protocol for the Swiss Context	
<i>Alexandre Debant</i>	7
A trust/distrust theoretical framework for internet voting	
<i>David Dueñas-Cid</i>	7
Beyond voting: tool building for better democracy	
<i>Bryan Ford</i>	8
Demo session: Coercion-resistant E-voting with in-person credentialing	
<i>Bryan Ford</i>	8
Which roadmap for a Software Development Kit suitable for Internet Voting?	
<i>Pierrick Gaudry</i>	9
Some thoughts about security definitions	
<i>Kristian Gjøsteen</i>	9
The Unstated Problem	
<i>Thomas Haines</i>	10
Fuzzing E-Voting Protocol Implementations	
<i>Lucca Hirschi</i>	10
Common trust assumptions in internet voting – how comfortable are we with them?	
<i>Audhild Høgåsen</i>	10
Defining vote privacy ... again	
<i>Steve Kremer</i>	11
Usable coercion-resistant voting	
<i>Oksana Kulyk</i>	11
E-Collecting in Switzerland	
<i>Reto König</i>	12
Verifiable Tally-Hiding E-Voting and zkSNARKs for Ballot Validity	
<i>Ralf Küsters</i>	12
Trusting implementations	
<i>Vincent Laporte</i>	12
Election Audits: There's More We Can Do	
<i>Jennifer Morrell</i>	13

Post-Quantum Proof of Shuffle	
<i>Rafieh Mosaheb</i>	13
Composing (provable) fine-grained definitions into an (understandable) E2E-verifiability definition	
<i>Florian Moser</i>	13
GI Elections: Individual Verification for POLYAS	
<i>Florian Moser</i>	14
Verifying cryptographic software in Lean	
<i>Michael Naehrig</i>	14
From individual verifiability to universally verified	
<i>Peter Y. A. Ryan</i>	14
Sign-Then-Encrypt	
<i>Peter Rønne</i>	15
Boardroom voting with their smartphones: Experiencing Verifiability	
<i>Kazuo Sako</i>	15
Some Usability Issues in Coercion-resistant and E2E Solutions	
<i>Roberto Samarone Araujo</i>	16
Selene Voter Experience	
<i>Steve Schneider</i>	16
Voting Channel Security	
<i>Carsten Schürmann</i>	16
Quantum-Safe Electronic Voting: Status and Challenges	
<i>Tjerdand Silde</i>	17
Electronic Voting in Brazil: How E2E-V might work there (a work in progress)	
<i>Marcos Simplicio</i>	17
Colorado IRV RLAs	
<i>Vanessa Teague</i>	17
Work in Progress: SoK on usable verifiable electronic voting	
<i>Melanie Volkamer</i>	18
Audit (Process) Efficiency for Plurality Elections	
<i>Poorvi Vora</i>	18

Working groups

Evidence-Based Elections (commitments)	
<i>Josh Benaloh</i>	18
Coercion Resistance	
<i>Constantin Catalin Dragan</i>	19
Methods for Trustworthy Randomness Generation	
<i>Jennifer Morrell and Peter Rønne</i>	19
Post-Quantum Cryptography and E-Voting	
<i>Rafieh Mosaheb and Peter Rønne</i>	20

Education and Outreach
Florian Moser 20

Software Development Kit for E-Voting
Florian Moser 21

Usability
Peter Rønne 21

Participants 22

3 Overview of Talks

3.1 My Journey to Commitment

Josh Benaloh (Microsoft Research – Redmond, US)

License © Creative Commons BY 4.0 International license
© Josh Benaloh

Most verifiable election systems work by encrypting votes. However, replacing encryptions with commitments for in-person voting greatly simplifies the process by removing the need for keys and trustees. Using commitments also substantially reduces the size and complexity of election records – greatly simplifying the burdens on verifiers.

Sparse commitments achieve all of these benefits without substantially altering basic integrity and privacy properties or the underlying assumptions. Dense commitments can achieve ever greater efficiencies as well as everlasting privacy of votes at a cost of making the integrity of the election dependent upon a computation assumption – such as the difficulty of computing discrete logarithms at the time of the election.

3.2 Risk-limiting Audits for Complex Elections: Progress and What’s Next

Michelle Blom (The University of Melbourne, AU), Alexander Ek (KU Leuven, BE), and Vanessa Teague (Australian National University – Acton, AU)

License © Creative Commons BY 4.0 International license
© Michelle Blom, Alexander Ek, and Vanessa Teague

In this talk we highlight the progress that has been made in the development of risk-limiting audits (RLAs) for complex elections, and the extent to which they have been adopted by practitioners. We discuss current challenges that are limiting further progress on RLAs for preferential and proportional elections, and propose several future directions for this research.

3.3 Breaking and fixing CHVote, a protocol for the Swiss context

Véronique Cortier (LORIA, CNRS – Nancy, FR)

License © Creative Commons BY 4.0 International license
© Véronique Cortier

Joint work of Véronique Cortier, Alexandre Debant, Pierrick Gaudry
Main reference Véronique Cortier, Alexandre Debant, Pierrick Gaudry: “Breaking verifiability and vote privacy in CHVote”, IACR Cryptol. ePrint Arch., p. 80, 2025.
URL <https://eprint.iacr.org/2025/080>

CHVote is one of the main electronic voting systems developed in the context of political elections in Switzerland, where the regulation requires a specific setting and specific trust assumptions. In this setting, voters must receive return codes to check that their device did not modify their vote. CHVote is based on oblivious transfer to achieve this functionality. We have shown that actually, CHVote fails to achieve vote secrecy and individual verifiability (here, recorded-as-intended), as soon as one of the online components is dishonest, contradicting the security claims of CHVote. The attacks have been presented at Esorics’25. An open question is to enhance the oblivious transfer scheme in order to retrieve privacy (and verifiability).

3.4 A Practical and Fully Distributed E-Voting Protocol for the Swiss Context

Alexandre Debant (INRIA – Villers-lès-Nancy, FR)

License © Creative Commons BY 4.0 International license
© Alexandre Debant

Joint work of Véronique Cortier, Alexandre Debant, Olivier Esseiva, Pierrick Gaudry, Audhild Høgåsen, Chiara Spadafora

Main reference Véronique Cortier, Alexandre Debant, Olivier Esseiva, Pierrick Gaudry, Audhild Høgåsen, Chiara Spadafora: “A Practical and Fully Distributed E-Voting Protocol for the Swiss Context”, IACR Cryptol. ePrint Arch., p. 1625, 2025.

URL <https://eprint.iacr.org/2025/1625>

Internet voting in Switzerland for political elections is strongly regulated by the Federal Chancellery (FCh). It puts a great emphasis on the individual verifiability: security against a corrupted voting device is ensured via return codes, sent by postal mail. For a long time, the FCh was accepting to trust an offline component to set up data and in particular the voting material. Today, the FCh aims at removing this strong trust assumption. We propose a protocol that abides by this new will. At the heart of our system lies a setup phase where several parties create the voting material in a distributed way, while allowing one of the parties to remain offline during the voting phase. A complication arises from the fact that the voting material has to be printed, sent by postal mail, and then used by the voter to perform several operations that are critical for security. Usability constraints are taken into account in our design, both in terms of computation complexity (linear setup and tally) and in terms of user experience (we ask the voter to type a high-entropy string only once). The security of our scheme is proved in a symbolic setting, using the ProVerif prover, for various corruption scenarios, demonstrating that it fulfills the Chancellery’s requirements and sometimes goes slightly beyond them

3.5 A trust/distrust theoretical framework for internet voting

David Dueñas-Cid (Kozłowski University, PL)

License © Creative Commons BY 4.0 International license
© David Dueñas-Cid

Trust is a complex concept that relies on numerous factors. Technology adds complexity to its understanding by providing a different sociotechnical context; on the one hand, the technical properties of technology contribute to the perception of trustworthiness in systems, but on the other, societal determinants affect individual perceptions with elements that extend beyond the systems’ technicalities. Electoral technologies do not simplify this equation. Other aspects, such as power distribution, result acceptance, political interests, or complex management, put even more pressure on trust-related issues. This presentation proposes a theoretical framework to structure the concepts of trust and distrust within the field, facilitating further research on the topic.

3.6 Beyond voting: tool building for better democracy

Bryan Ford (EPFL Lausanne, CH)

License  Creative Commons BY 4.0 International license
 Bryan Ford

In this talk I propose the need to broaden our problem scope from “end-to-end voting” to “end-to-end democracy”, incorporating not just the process of casting, collecting, and counting votes, but also the process of democratic agenda-setting and decision-making. In particular, there are promising scalable processes such as random-sampled citizens’ assemblies or mini-publics, and liquid democracy, that show the potential to enable people to participate more directly at large scale, gather the information and diverse perspectives to make better decisions in everyone’s broader interests, and leverage accountable expertise to perform deep analysis of complex issues. But to make these processes work at scale in borderless, permissionless deployment settings, we need the secure and privacy-preserving foundation of proof of personhood, and that foundation needs to be coercion resistant in order to ensure that participants are representing their own interests rather than someone else’s. I see promise that this challenge is solvable in principle because at least in-person proof of personhood (pseudonym parties) are composable with coercion-resistance mechanisms based on fake credentials and in-person credentialing as developed in the TRIP/Votegral system.

3.7 Demo session: Coercion-resistant E-voting with in-person credentialing

Bryan Ford (EPFL Lausanne, CH)

License  Creative Commons BY 4.0 International license
 Bryan Ford

Joint work of Louis-Henri Merino, Simone Colombo, Rene Reyes, Alaleh Azhir, Shailesh Mishra, Pasindu Tennage, Mohammad Amin Raeisi, Haoqian Zhang, Jeff R. Allen, Bernhard Tellenbach, Vero Estrada-Galiñanes, Bryan Ford

Main reference Louis-Henri Merino, Simone Colombo, Rene Reyes, Alaleh Azhir, Shailesh Mishra, Pasindu Tennage, Mohammad Amin Raeisi, Haoqian Zhang, Jeff R. Allen, Bernhard Tellenbach, Vero Estrada-Galiñanes, Bryan Ford: “TRIP: Coercion-resistant Registration for E-Voting with Verifiability and Usability in Votegral”, in Proc. of the ACM SIGOPS 31st Symposium on Operating Systems Principles, SOSP ’25, p. 834–874, Association for Computing Machinery, 2025.

URL <https://doi.org/10.1145/3731569.3764837>

This talk briefly summarized the design and usability study results for TRIP, the coercion-resistant in-person registration process for end-to-end verifiable electronic voting in the Votegral system. The slides summarize the design and key usability study results, and the linked video is one of the three instructional videos we used to explain the system to participants in the usability study.

3.8 Which roadmap for a Software Development Kit suitable for Internet Voting?

Pierrick Gaudry (CNRS – Nancy, FR)

License  Creative Commons BY 4.0 International license
© Pierrick Gaudry

In this talk, we propose to explore the possibility to build a Software Development Kit (SDK) that could be the basis for constructing various Internet Voting systems. Several questions are to be answered: which features do we want to support? Which programming language? Which licence? How to organize the governance and the maintenance? And last but not least, can we build this SDK on top of existing SDK like ElectionGuard (for in-person voting) or Verificatum (for mixnets), or by extracting parts of existing full system implementations like CHVote, SwissPost, Helios, Belenios, Polyas, etc?

3.9 Some thoughts about security definitions

Kristian Gjølsteen (NTNU – Trondheim, NO)

License  Creative Commons BY 4.0 International license
© Kristian Gjølsteen

To make security results comparable, we would like to have one security definition that covers (almost) all systems, even paper-based systems and systems with more than one casting method. We also want to model all aspects of the system, including human behaviour/mistakes.

To define security, we first need to model the class of voting systems. The thing every system has in common is that there are voters that want to cast ballots, there should be a result, and there will be an adversary.

As usual, an experiment simulates the honest parties and interacts with an adversary. The adversary directs the system, including starting honest parties, scheduling message delivery, and directing honest voter operations (casting, verifying, reporting results, etc.) and signalling the next election phase (assumed to be synchronised). The adversary may also corrupt parties.

The experiment keeps track of the honest voters' beliefs about what ballots they have successfully cast. For integrity, our goal is two-fold: any accepted result is consistent with the honest voters' beliefs, and any result is unique. For accountability, parties that evaluate the result output evidence that anyone can evaluate to point to a party, and our goal is that any party pointed at is corrupt. For coercion resistance, the adversary defines two groups of voters and attempts to coerce all of them. One group resists coercion, the other does not. The adversary's goal is to distinguish the groups.

3.10 The Unstated Problem

Thomas Haines (Australian National University – Acton, AU)

License  Creative Commons BY 4.0 International license
© Thomas Haines

What do the vulnerabilities we see in deployed voting systems tell us about our definitions? It suggests that we constrain attacks on state too much in definitions and not enough in our protocols.

3.11 Fuzzing E-Voting Protocol Implementations

Lucca Hirschi (LORIA & INRIA – Villers-lès-Nancy, FR)

License  Creative Commons BY 4.0 International license
© Lucca Hirschi

Joint work of Max Ammann, Lucca Hirschi, Steve Kremer
Main reference Max Ammann, Lucca Hirschi, Steve Kremer: “DY Fuzzing: Formal Dolev-Yao Models Meet Cryptographic Protocol Fuzz Testing”, in Proc. of the IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024, pp. 1481–1499, IEEE, 2024.
URL <https://doi.org/10.1109/SP54263.2024.00096>

Today’s information society crucially relies on cryptographic protocols. These protocols leverage cryptographic primitives to ensure confidentiality, integrity, or other security goals. A widespread class of vulnerabilities plaguing such protocols is logical attacks, which exploit flawed protocol logic. We examine two formal model-based methods for preventing logical attacks in protocol designs and implementations.

First, we present automated formal verification methods based on Dolev-Yao (DY) models, which formally define and excel at finding such flaws on protocol design specifications. Nevertheless, these methods alone cannot secure protocol implementations, as bugs may introduce implementation-level logical attacks. We then present a recent research avenue that aims to integrate formal DY models with fuzz testing techniques to capture logical attacks in cryptographic protocol implementations.

Finally, we explore the idea of fuzzing e-voting protocols, possibly using DY fuzzing. DY fuzzing seems ideally placed to address the unique challenges posed by e-voting systems, which constitute particularly interesting targets for several reasons: they are based on many different sub-protocols run by different actors, they aim to achieve extremely strong security properties under strong threat models, and they are critical pieces of software which have not received much attention from the fuzzing community so far.

3.12 Common trust assumptions in internet voting – how comfortable are we with them?

Audhild Høgåsen (Schweizerische Post – Bern, CH)

License  Creative Commons BY 4.0 International license
© Audhild Høgåsen

We discuss how comfortable we are with some common trust assumptions for internet voting. Further, we discuss the following question: “Which do you consider preferable for security for a scheme being implemented today; a protocol which security is based on lattice assumptions

on a protocol which security is based on classical assumptions?” We also discuss questions like “Would you, in order to get coercion resistance, accept the risk that some voters are prevented from voting?” Finally, we ask ourselves: “If you must choose only 2 of the following, which do you choose?”: “Vote secrecy against malicious voter device”, “Individual verifiability (against malicious voter device)”, “Coercion resistance”, “Post-quantum security (on top of classical security, so assume some hybrid scheme)”.

3.13 Defining vote privacy ... again

Steve Kremer (INRIA – Villers-lès-Nancy, FR)

License  Creative Commons BY 4.0 International license
© Steve Kremer

Defining vote privacy is tricky. In this talk we review some classical game-based definitions, recall some known short-comings of these definitions, and show a few new problems. In particular, we demonstrate that one classical definition, generally believed too strong for some counting functions, may actually be too weak. We also observe that the BPriv definition, regarded as one of the most mature definitions, is too strong for protocols that chain votes (which may be needed to avoid a recent attack). By supposing an “extract function” (a hypothesis that already exists in BPriv) we propose a simple variant of a classical definition by Benaloh that seems to overcome the mentioned problems.

3.14 Usable coercion-resistant voting

Oksana Kulyk (IT University of Copenhagen, DK)

License  Creative Commons BY 4.0 International license
© Oksana Kulyk

Joint work of Christina Nissen, Tobias Hilt, Jurlind Budurushi, Melanie Volkamer, Oksana Kulyk
Main reference Christina Nissen, Tobias Hilt, Jurlind Budurushi, Melanie Volkamer, Oksana Kulyk: “Voting Under Pressure: Perceptions of Counter-Strategies in Internet Voting”, in Proc. of the Electronic Voting – 10th International Joint Conference, E-Vote-ID 2025, Nancy, France, October 1-3, 2025, Proceedings, Lecture Notes in Computer Science, Vol. 16028, pp. 158–174, Springer, 2025.
URL https://doi.org/10.1007/978-3-032-05036-6_10

The risks of voter coercion in online voting has lead to a body research proposing coercion-resistant solutions, that would allow a coerced voter to adopt a counter-strategy that manages to deceive the coercer while at the same time allowing the voter to vote according to their actual wishes. The usability of these solutions, and voters’ ability to successfully apply them in case of coercion, however, remains an open question. In this talk I present our ongoing research on the usability of different coercion-resistant strategies, concluding that while certain strategies can potentially be made usable, more research is needed to address their limitations.

3.15 E-Collecting in Switzerland

Reto König (Bern University of Applied Sciences, CH)

License  Creative Commons BY 4.0 International license
© Reto König

After a real attack on the collection process for votes on initiatives and referenda (“Unterschriftenbschiss”), Switzerland faces is urged to implement a more secure and thus trustworthy process on how to collect such votes – electronically. Even though the process seems to result in a “lightweight” e-voting protocol, the security requirements are quite hefty. In order to protect Swiss Citizens from losing all their privacy, the system is required to provide “unconditional” ever lasting participation privacy. Furthermore, the system needs to be sound regarding eligibility throughout the entire procedure which can last up to 18 month. In seeking help from our fellow e-voting experts, we propose a system based on set-membership ZKPs with a single-use rate limiter. But will it be enough?

3.16 Verifiable Tally-Hiding E-Voting and zkSNARKs for Ballot Validity

Ralf Küsters (Universität Stuttgart, DE)

License  Creative Commons BY 4.0 International license
© Ralf Küsters

Modern electronic voting systems (e-voting systems) are designed to provide not only vote privacy but also (end-to-end) verifiability. Several verifiable e-voting systems have been proposed in the literature. Most existing systems reveal not just the voting result but also the full tally, consisting of the exact number of votes per candidate or even all single votes. There are several situations where this is undesirable. So-called tally-hiding systems solve this problem. In this talk, several forms of tally-hiding are discussed and two systems, namely Ordinos and Kryvos, are presented to achieve verifiable yet tally-hiding e-voting, including a post-quantum secure system. Moreover, a new approach for ensuring ballot validity is presented based on zkSNARKs.

3.17 Trusting implementations

Vincent Laporte (LORIA & INRIA – Villers-lès-Nancy, FR)

License  Creative Commons BY 4.0 International license
© Vincent Laporte

The overall security of a voting infrastructure relies, to some extend, on the quality of software implementations. This talk reminds some desirable features of an implementation and explores whether voting systems could benefit from high-assurance implementations. It also presents the Jasmin system for high-assurance cryptography.

3.18 Election Audits: There's More We Can Do

Jennifer Morrell (University of Minnesota – Minneapolis, US)

License © Creative Commons BY 4.0 International license
© Jennifer Morrell

Election audits are vital to accountability and public confidence, yet their methods have become increasingly politicized. This presentation advances nine standards for post-election tabulation audits – covering independence, transparency, reconciliation, and voter privacy – and connects them to trusted frameworks in other policy domains. Drawing on experimental research, we show that process-focused visual audit summaries significantly increase voter confidence, especially among independents. We translate these findings into practical templates to help election officials communicate results more effectively. The presentation concludes by exploring automation and AI-assisted tools that can enhance audit quality, transparency, and accountability to voters.

3.19 Post-Quantum Proof of Shuffle

Rafieh Mosaheb (University of Luxembourg, LU)

License © Creative Commons BY 4.0 International license
© Rafieh Mosaheb
Joint work of Thomas Haines, Rafieh Mosaheb, Johannes Müller
Main reference Thomas Haines, Rafieh Mosaheb, Johannes Müller, Reetika: “Zero-Knowledge Proofs from Learning Parity with Noise: Optimization, Verification, and Application”, in Proc. of the 38th IEEE Computer Security Foundations Symposium, CSF 2025, Santa Cruz, CA, USA, June 16-20, 2025, pp. 441–456, IEEE, 2025.
URL <https://doi.org/10.1109/CSF64896.2025.00010>

In this talk, we present the design of a post-quantum e-voting system based on the hardness of the Learning Parity with Noise (LPN) problem. We begin by motivating our choice of this approach and highlighting the importance of diversifying the underlying post-quantum assumptions in the context of e-voting. Next, we describe the phases of our proposed system and the cryptographic primitives employed. We then detail the construction of our code-based proof of shuffle. The talk concludes with a discussion of the current limitation of this proof of shuffle, specifically, its inability to distribute trust among multiple talliers, requiring reliance on a single trusted tallier to preserve ballot privacy. We are exploring directions to overcome this limitation, either by developing a code-based proof of shuffle that supports distributed trust or by rethinking the structure of the e-voting phases to achieve this goal.

3.20 Composing (provable) fine-grained definitions into an (understandable) E2E-verifiability definition

Florian Moser (famoser GmbH – Allschwil, CH)

License © Creative Commons BY 4.0 International license
© Florian Moser

For an intuitive understanding of the overall security guarantees of an internet voting system, a single and short high-level definition such as E2E-Verifiability is useful (e.g., roughly, all ballots are correctly tallied exactly as cast). In contrast, when proving the security of such

a system, or when specifying precise behaviour of some part of the system, fine-grained definitions are more practical. However, proving certain fine-grained properties does not necessarily imply that the targeted high-level property is indeed fulfilled, too.

In this work, we aim to bring these two approaches together. We propose precise fine-granular definitions, inspired by usual notions in literature. Then, we compose these fine-granular definitions and observe the exact E2E-Verifiability notion reached. Finally, we propose avenues for future research, which includes most crucially a proof of our claims.

3.21 GI Elections: Individual Verification for POLYAS

Florian Moser (famoser GmbH – Allschwil, CH)

License  Creative Commons BY 4.0 International license
 © Florian Moser
URL <https://github.com/famoser/polyas-verification>

When voting over the internet, the voter needs to be able to verify that their vote was cast correctly. This avoids trusting their voting device fully. This project implements voter-side verification (also called individual verification) for POLYAS 3.0 (Version 1.3.2, 31 July 2023), following the second device spec (Version 1.2-SNAPSHOT, 22. September 2024).

This project has been developed for the Gesellschaft für Informatik (German) and has been supported by the Université de Lorraine, CNRS, Inria, and LORIA (Nancy, France). In collaboration with other researchers, a scientific publication documents the experience of implementing this verifier (and others), and formulates recommendation for similar projects: <https://inria.hal.science/hal-04663997>.

3.22 Verifying cryptographic software in Lean

Michael Naehrig (Microsoft – Redmond, US)

License  Creative Commons BY 4.0 International license
 © Michael Naehrig

In a collaborative effort, we are modernizing Microsoft’s cryptography library SymCrypt by transitioning implementations of cryptographic primitives to Rust. These Rust implementations are then formally verified using the Aeneas tool chain to produce an equivalent Lean representation that is shown to be functionally correct according to the primitive’s specification. This talk describes the methodology and discusses the reasons for this approach.

3.23 From individual verifiability to universally verified

Peter Y. A. Ryan (University of Luxembourg – Esch-sur-Alzette, LU)

License  Creative Commons BY 4.0 International license
 © Peter Y. A. Ryan

For End-to-end verifiable voting systems a key element of the assurance argument is that the vote is *cast as intended*: voters should be confident that their intended vote was encrypted. This often takes the form of an audit of the ballot: the encryption of the vote is opened to reveal the plaintext. It is often thought that an audited ballot cannot then be cast as this

would violate receipt-freeness: auditing typically requires revealing the randomisation used for the encryption. This leads to the use of some form of cut-and-choose protocol between the voter and the voting device, for example, Benaloh challenges.

In this talk we argue that a cut-and-choose procedure may not in fact be necessary: if an audited ballot is re-encrypted before being posted then receipt-freeness can be preserved. The voter still needs a way to identify their ballot on the BB, but this can be achieved, for example, using suitable malleable signatures that remain valid under re-encryption of the signed text. The assurance provided by being able to cast an audited ballot is far higher than that possible via cut-and-choose techniques.

A further observation is that if this approach is used in conjunction with a scheme that involves pre-commitment to encryptions of the votes these can be independently pre-audited. This removes the need for voters to perform their own ballot audits. Given that such audits are generally regarded as burdensome this will contribute to improved usability, and means that the assurance in the outcome is less reliant on a good proportion of voters performing the checks diligently.

3.24 Sign-Then-Encrypt

Peter Rønne (University of Luxembourg, LU)

License  Creative Commons BY 4.0 International license
© Peter Rønne

In many voting schemes the plaintext vote choice is being encrypted with zero-knowledge proofs of ballot correctness to form a ballot which is then digitally signed. This digital signature facilitates voter eligibility checks and prevents ballot stuffing. In this talk we suggest improving cast-as-intended verifiability by instead signing the plaintext vote directly using one device after which a second device encrypts the choice and provides a zero-knowledge proof that it has seen a valid signature of an allowed vote choice. To maliciously change a vote undetected, an attacker would need to corrupt both devices. This is especially interesting if the voting system can rely on an already existing electronic ID infrastructure and corresponding secure signing devices. This generalises ideas from schemes using blind, re-randomisable and linearly-homomorphic signatures.

3.25 Boardroom voting with their smartphones: Experiencing Verifiability

Kazue Sako (Waseda University – Tokyo, JP)

License  Creative Commons BY 4.0 International license
© Kazue Sako

In my talk, I will call for proposals in a specific, real-life voting scenario. We had been discussing about in-person voting where casting vote takes in a shared computer at a polling place, and remote voting where people cast their vote through their individual computers/smartphones. The scenario I propose today is a mixture – it will be in-person, in a boardroom where people in the room are presumably authenticated, but 300 voters will be using their own smartphones to do the voting. The primary purpose is to have a shorter voting/tallying period, but this aims to provide voters the value of verifiability that is hard to achieve with plain paper ballots.

3.26 Some Usability Issues in Coercion-resistant and E2E Solutions

Roberto Samarone Araujo (Federal University of Pará – Belém, BR)

License  Creative Commons BY 4.0 International license
© Roberto Samarone Araujo

Coercion resistance and end-to-end verifiability represent two of the most promising approaches for enhancing the security and transparency of modern elections. Despite their potential, these technologies still face significant challenges that hinder their use in real-world electoral settings. In this talk, we examine key usability issues associated with these approaches and discuss two potential strategies that could mitigate such limitations.

3.27 Selene Voter Experience

Steve Schneider (University of Surrey – Guildford, GB)

License  Creative Commons BY 4.0 International license
© Steve Schneider

Joint work of Mohammed Alsadi, Steve Schneider

Main reference Robert Krimmer, Melanie Volkamer, Bernhard Beckert, Ardita Driza Maurer, David Dueñas-Cid, Stéphane Glondou, Iuliia Krivonosova, Oksana Kulyk, Ralf Küsters, Beata Martin-Rozumilowicz, Peter Rønne, Mihkel Solvak, Oliver Spycher: “Fifth International Joint Conference on Electronic Voting – E-Vote-ID 2020 – Taltech Press Proceedings”, 2020.

URL https://www.researchgate.net/publication/344471755_Fifth_International_Joint_Conference_on_Electronic_Voting_-_E-Vote-ID_2020_-_Taltech_Press_Proceedings#page=302

Selene is a verifiable voting protocol proposed by Ryan, Roenne and Iovino. It uses trackers to provide a way of enabling voters to check their vote in plaintext at the end of the election. It can be added to existing voting schemes to add verifiability functionality.

This talk described a trial “in the wild” with a commercial internet voting provider, whereby their existing (non-verifiable) system was augmented with additional functionality based on Selene verifiability (though with stronger trust assumptions) to give voters the opportunity to check their votes. This was used in two real online elections in order to explore voters’ experience of verifiability and their opinions of it, by means of a questionnaire. The majority of the respondents found the system easy to use and considered that it maintained their vote privacy. They appreciated the opportunity to verify their vote and felt it gave them confidence in the system, but they were split on seeing the point of verifiability. This leaves open the question of how to engender trust in electronic voting systems and to what extent this needs to relate to understanding.

3.28 Voting Channel Security

Carsten Schürmann (IT University of Copenhagen, DK)

License  Creative Commons BY 4.0 International license
© Carsten Schürmann

Elections are designed to achieve a peaceful resolution to a civic battle about power and governance. Implied is that elections should ensure that all voters – supporting winners and losers alike – accept the election outcome and develop collective trust. In my talk, I argue that if an election offers multiple voting channels, their interactions might be subtle, and they can have negative effects on the collective trust. Politicization driven by political parties, activism, and foreign influence of existing voting channels, malicious or accidental, can trigger security violations that puts this collective trust at risk.

3.29 Quantum-Safe Electronic Voting: Status and Challenges

Tjerand Silde (NTNU – Trondheim, NO)

License © Creative Commons BY 4.0 International license
© Tjerand Silde

Quantum computers are coming and will break the (long-term) privacy and (online) integrity of electronic voting schemes; hence, we need to build new schemes from quantum-safe assumptions. This talk will overview the quantum threat, lattice-based cryptography, and our current state-of-the-art schemes.

3.30 Electronic Voting in Brazil: How E2E-V might work there (a work in progress)

Marcos Simplicio (University São Paulo, BR)

License © Creative Commons BY 4.0 International license
© Marcos Simplicio

The talk briefly shows how electronic voting in Brazil currently works: it is based on Direct Recording Electronic (DRE) machines empowered by Trust Platform Modules (TPM) for protecting its underlying software – so they are called “T-DRE” machines. To improve the system’s transparency, it is possible to integrate end-to-end verifiability (E2E-V) features into the process, thus giving voters the ability to check if their individual votes were cast-as-intended, recorded-as-cast, and tallied-as-recorded. The presentation shows the proposed procedures, which include: tracking codes generated by voting machines via cryptographic commitments (e.g., using Pedersen commitments for ever-lasting privacy); Benaloh challenges on those commitments, done under the supervision of auditors, with simulated votes, to preserve vote privacy; signature of cast votes, aiming to facilitate dispute resolution if a tracking code does not appear in the tally; and publication of all relevant records. It also discusses the confidence level of the election results, depending on the number of challenges performed and the difference in the number of votes received by each candidate.

3.31 Colorado IRV RLAs

Vanessa Teague (Australian National University – Acton, AU)

License © Creative Commons BY 4.0 International license
© Vanessa Teague
Joint work of Vanessa Teague, Michelle Blom, Andrew Conway
URL <https://github.com/democracydevelopers>

We’ll talk through the educational materials, and possibly also demo the code, at <https://github.com/democracydevelopers>.

This extends Colorado’s Risk Limiting Audit code to IRV, based on the RAIRE system (Blom, Stuckey, Teague, <https://arxiv.org/abs/1903.08804>).

The code will run in Colorado’s RLA this November.

3.32 Work in Progress: SoK on usable verifiable electronic voting

Melanie Volkamer (KIT – Karlsruher Institut für Technologie, DE)

License  Creative Commons BY 4.0 International license
© Melanie Volkamer

Based on our literature review on usability studies for verifiable electronic voting systems, we identified more than 20 factors in which the study setting differs, such as the type of mechanism studied, whether it was a remote or a lab study, what instructions they got, and what type of manipulations were studied. Given the number of papers we found, we conclude that it is difficult to compare the results. We learned that the most challenging part is to measure the effectiveness of the verifiability mechanism: One can observe e.g. by eye tracking whether voters go for the verification step at all but does this mean they would notice the manipulation and if yes would they report it properly.

3.33 Audit (Process) Efficiency for Plurality Elections

Poorvi Vora (George Washington University – Washington, DC, US)

License  Creative Commons BY 4.0 International license
© Poorvi Vora

We provide an example where a mathematical model which better represents the physical audit process greatly reduces the expected number of ballots drawn in a ballot polling RLA, and thus reduces the number of person-hours required for the audit. We hence argue that better models of the true audit process on the ground will generally lead to (different) statistical tests representing “better” real audits.

4 Working groups

This section presents abstracts of selected breakout groups from the Dagstuhl Seminar.

4.1 Evidence-Based Elections (commitments)

Josh Benaloh (Microsoft Research – Redmond, US)

License  Creative Commons BY 4.0 International license
© Josh Benaloh

This discussion focused primarily on using cryptographic commitments rather than encryption for verifiable voting systems. This technique is most useful for in-person voting where it enables the complete elimination of cryptographic keys and the trustees who would generate and manage the keys.

The discussion was generally quite favorable, although an important point was raised that backup can create privacy risks that are not present with the prior approach of using encryption to protect the privacy of votes. With the encryption approach, ballots are independently encrypted and no intermediate tally data is retained. However, the commitment approach effectively maintains a running tally after each new ballot is cast,

and a full sequence of intermediate tallies reveals the contents of each individual vote. The concern is, however, mitigated by the fact the the privacy risk is no greater than that posed by traditional ballot tallying systems which also maintain running tallies.

4.2 Coercion Resistance

Constantin Catalin Dragan (University of Surrey – Guildford, GB)

License  Creative Commons BY 4.0 International license
© Constantin Catalin Dragan

Coercion has typically been viewed as “persuading someone to do something by using force or threats” (Oxford dictionary). The voting community has proposed some solutions for when the voter can leave the coercive environment (e.g., fake credentials in JCJ/Civitas or re-voting in Belenios) or under extreme coercion – can never leave the environment and can only annul their vote.

One of biggest challenges is to define what makes acceptable manipulation versus actual coercion. The voting community aims to technically and ethically support governments in tackling this challenge. One potential context is we’ve seen a recent increase in electoral influence via social media that can have drastic consequences on elections, for example the cancelled 2024 Romanian presidential election or the Cambridge Analytica scandal on influencing the 2016 US presidential elections. Another big challenge is on understanding the trade-offs between problems and solutions, with a focus on designing relevant baselines for comparison (e.g., failure rates, voter expectations, and real-world testing). It is exceptionally hard to test any developed solutions under high-stress conditions to simulate the real-world voting environment.

4.3 Methods for Trustworthy Randomness Generation

Jennifer Morrell (University of Minnesota – Minneapolis, US) and Peter Rønne (University of Luxembourg, LU)

License  Creative Commons BY 4.0 International license
© Jennifer Morrell and Peter Rønne

This working group discussed methods for trustworthy randomness generation used in audits of elections. The current most-used method consists in rolling dice for seed to a pseudo-random function, where the ceremony is public or recorded and broadcasted. Different attacks were discussed such as deliberate biasing (or predicting) of the sample, but also the accusation that the sample is biased or predictable (when it isn’t). New constructions and protocols were discussed, especially with active contribution from people who wish to participate in the ceremony.

4.4 Post-Quantum Cryptography and E-Voting

Rafieh Mosaheb (University of Luxembourg, LU) and Peter Rønne (University of Luxembourg, LU)

License  Creative Commons BY 4.0 International license
© Rafieh Mosaheb and Peter Rønne

Currently, there is a large ongoing transformation to post-quantum cryptography due to potential threats from upcoming quantum computers against classical asymmetric cryptography.

This breakout group discussed the current status of quantum-safe e-voting, threats and solutions including partial solutions such as voting with perfect privacy that protects against store-now-decrypt-later attacks. In particular, possible constructions for hybrid verifiable encryption were discussed as a way to protect election privacy by combining classical and post-quantum cryptography, ensuring security even if a PQ scheme is later found to be flawed, while still providing verifiability against quantum attackers.

4.5 Education and Outreach

Florian Moser (famoser GmbH – Allschwil, CH)

License  Creative Commons BY 4.0 International license
© Florian Moser

The participants identified different target public that need to be reached with education and outreach projects. The voters need to know how to use the system, but additionally need to be able to react to unforeseen states (e.g., a failed verification check), and to battle misinformation (e.g., claims about frauds of the election). The providers need to be able to continuously improve their system, given their respective resources and expertise. Last but not least, also non-experts that interact deeply with these systems need to be empowered to take informed decisions, which includes the election organizers, activists and observers.

However, many aspects of verifiable elections are hard to communicate. First, many results are non-intuitive, but nonetheless correct. For example, risk-limiting audits in elections with large margins need to audit very few ballots to achieve their guarantees (e.g. in the Kenyan elections 2017, 200 ballots for the whole country would have been enough for 99.9% of confidence). Also, the cryptography used to anonymize ballots before decryption are not well known outside of the internet voting domain, and may be unintuitive even for otherwise experts in IT security. The complexity and the “rough edges” of the current state-of-the-art systems may lead to distrust, or even the choice of an insecure but perceived simpler system.

The participants also discussed how to concretely move forward. Some of the measures discussed are to audit & update Wikipedia articles concerning verifiable elections, and to approach the development of an SDK for system providers (as also discussed in another session in this seminar).

4.6 Software Development Kit for E-Voting

Florian Moser (famoser GmbH – Allschwil, CH)

License  Creative Commons BY 4.0 International license
© Florian Moser

The working groups had a session discussing the development of a Software Development Kit for voting. The targets for the SDK should be verifiable elections, both for remote voting and in-person voting. While previous work exists, these SDKs usually cover only part of the cryptography used in verifiable elections, or are adapted to the needs of a specific system or vendor.

Some of the challenges discussed were the intended usability both for a target audience without a deep cryptographic background, while allowing customizability for cryptographers for special use cases. A basis for the the initial SDK would be public key encryption, commitments, signatures and zero-knowledge proofs, as well as procedures for distributed key generation, distributed decryption, verifiable shuffles and homomorphic tallies. Further, it should provide building blocks for public bulletin boards and tamper-proof logging. Documentation should also include instructions for auditors of elections.

The governance aspects were also discussed, especially with a view towards long-term stability of the SDK by contacting and inviting stake-holders, from industry, administration and vendors, early in the process. The library does not need to start afresh, but can profit from existing libraries and their experiences, especially ElectionGuard.

4.7 Usability

Peter Rønne (University of Luxembourg, LU)

License  Creative Commons BY 4.0 International license
© Peter Rønne

The discussion in this working group was usability, user experience and trust. Usability encompasses several key aspects: the vote casting process, voter verifiability – including verifying that votes are both cast-as-intended and stored-as0cast but sometimes also more advanced mechanisms, e.g., ensuring resistance towards coercion. A problem with the current usability and user experience research is that it exhibits a high variability in study design, which demonstrates a need for clearer guidelines that address different manipulation types and the reporting of detected issues in the voting process especially for verifiability. Building trust in elections is a deeply interdisciplinary challenge that demands collaborative research. Future studies should explore how different types of information affect trust, particularly messages that encourage voters to verify or report issues. Additionally, while existing research focuses mainly on voters, more attention should be given to poll workers. Finally, two specific research gaps persist: effectively communicating new verification capabilities to voters accustomed to “black box” systems, and understanding the long-term effects of voter verifiability.

Participants

- Josh Benaloh
Microsoft Research –
Redmond, US
- Michelle Blom
The University of Melbourne, AU
- Véronique Cortier
LORIA, CNRS – Nancy, FR
- Alexandre Debant
INRIA – Villers-lès-Nancy, FR
- Thi Van Thao Doan
University of Louvain, BE
- Constantin Catalin Dragan
University of Surrey –
Guildford, GB
- David Dueñas-Cid
Kozminski University, PL
- Alexander Ek
KU Leuven, BE
- Bryan Ford
EPFL Lausanne, CH
- Pierrick Gaudry
CNRS – Nancy, FR
- Kristian Gjøsteen
NTNU – Trondheim, NO
- Thomas Haines
Australian National University –
Acton, AU
- J. Alex Halderman
University of Michigan –
Ann Arbor, US
- Lucca Hirschi
LORIA & INRIA –
Villers-lès-Nancy, FR
- Audhild Høgåsen
Schweizerische Post – Bern, CH
- Reto König
Bern University of Applied
Sciences, CH
- Steve Kremer
INRIA – Villers-lès-Nancy, FR
- Ralf Küsters
Universität Stuttgart, DE
- Oksana Kulyk
IT University of
Copenhagen, DK
- Vincent Laporte
LORIA & INRIA –
Villers-lès-Nancy, FR
- Karola Marky
Ruhr-Universität Bochum, DE
- Jennifer Morrell
University of Minnesota –
Minneapolis, US
- Rafieh Mosaheb
University of Luxembourg, LU
- Florian Moser
famoser GmbH – Allschwil, CH
- Michael Naehrig
Microsoft – Redmond, US
- Olivier Pereira
University of Louvain, BE
- Thomas Peters
University of Louvain, BE
- Bart Preneel
KU Leuven, BE
- Sylvain Ruhault
ANSSI – Paris, FR
- Mark D. Ryan
University of Birmingham, GB
- Peter Y. A. Ryan
University of Luxembourg –
Esch-sur-Alzette, LU
- Peter Rønne
University of Luxembourg, LU
- Kazue Sako
Waseda University – Tokyo, JP
- Roberto Samarone Araujo
Federal University of Pará –
Belém, BR
- Steve Schneider
University of Surrey –
Guildford, GB
- Carsten Schürmann
IT University of
Copenhagen, DK
- Tjerand Silde
NTNU – Trondheim, NO
- Marcos Simplicio
University São Paulo, BR
- Vanessa Teague
Australian National University –
Acton, AU
- Melanie Volkamer
KIT – Karlsruher Institut für
Technologie, DE
- Poorvi Vora
George Washington University –
Washington, DC, US



Creativity, GenAI, and Software Development: A Future Together

Victoria Jackson^{*1}, Rafael Prikladnicki^{†2}, Daniel Russo^{†3},
Margaret-Anne Storey^{†4}, and André van der Hoek^{†5}

1 University of Southampton, GB. v.jackson@soton.ac.uk

2 Pontifícia Universidade Católica do Rio Grande do Sul, BR.
rafael.prikladnicki@pucrs.br

3 Aalborg University, DK. daniel.russo@cs.aau.dk

4 University of Victoria, CA. mstorey@uvic.ca

5 University of California Irvine, US. andre@ics.uci.edu

Abstract

This report introduces the goals of the Dagstuhl Perspectives Workshop 25412 “Creativity, GenAI, and Software Development: A Future Together”, provides an overview of the workshop and how it was conducted, details the activities that took place to identify significant future themes concerning research in creativity, GenAI, and software development, and foreshadows key outcomes that will be included in the associated Dagstuhl Manifesto.

Seminar October 5–10, 2025 – <https://www.dagstuhl.de/25412>

2012 ACM Subject Classification Software and its engineering → Software development techniques; Human-centered computing → Collaborative and social computing

Keywords and phrases Creativity, Generative AI, Software Development, Dagstuhl Perspectives Workshop

Digital Object Identifier 10.4230/DagRep.15.10.23

1 Executive Summary

Victoria Jackson (University of Southampton, GB, v.jackson@soton.ac.uk)

Rafael Prikladnicki (Pontifícia Universidade Católica do Rio Grande do Sul, BR, rafael.prikladnicki@pucrs.br)

Daniel Russo (Aalborg University, DK, daniel.russo@cs.aau.dk)

Margaret-Anne Storey (University of Victoria, CA mstorey@uvic.ca)

André van der Hoek (University of California, Irvine, US, andre@ics.uci.edu)

License © Creative Commons BY 4.0 International license

© Victoria Jackson, Rafael Prikladnicki, Daniel Russo, Margaret-Anne Storey, and André van der Hoek

This Dagstuhl Perspectives Workshop 25412 “Creativity, GenAI, and Software Development: A Future Together” brought together leading researchers and thought leaders from industry to collectively examine the evolving relationship between creativity, Generative AI (GenAI), and software development, and to articulate a forward-looking research agenda at this critical inflection point for the field.

Although much of the discourse, especially in industry, has focused on productivity gains from GenAI, this workshop deliberately shifted attention toward creativity as a central dimension of software development work.

* Editorial Assistant / Collector

† Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Creativity, GenAI, and Software Development: A Future Together, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 23–36

Editors: Victoria Jackson, Rafael Prikladnicki, Daniel Russo, Margaret-Anne Storey, and André van der Hoek



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

During five days of highly interactive and trust-building activities, participants explored how creativity manifests in software development, how it is shaped by tools, teams, and organizational contexts, and how GenAI can both amplify and erode creative practice. Rather than treating creativity as an abstract or purely individual trait, discussions emphasized its situated, socio-technical, and collective nature, spanning problem framing, exploration, sensemaking, and decision-making throughout the software lifecycle.

A key insight emerging from the workshop is that GenAI, in disrupting software development by reconfiguring who does what and in changing how value is created and assessed, creativity is simultaneously more important than ever and seriously challenged. Participants surfaced optimistic futures, in which GenAI expands creative reach and lowers barriers to exploration, and more cautionary trajectories, where over-automation risks flattening diversity of thought, narrowing solution spaces, or displacing reflective practice. These tensions highlight the need for research that goes beyond tool performance to examine creative quality, meaning, values, and long-term impacts on people and organizations.

The workshop culminated in the identification of five interrelated research directions that form the basis of the forthcoming Dagstuhl Manifesto, provisionally titled “Creativity is the New Productivity.” These directions call for: (1) developing a critical perspective on GenAI and creativity, (2) deeper characterization of creativity in software development; (3) stronger evidence of how creative work generates value; (4) exploration of the design space for GenAI-driven tools and processes that responsibly amplify creativity; and (5) synthesis across fragmented understandings, perspectives, and emergent publications to effectively promote knowledge sharing and generate impact.

Beyond the Manifesto itself, the workshop fostered a shared language, durable collaborations, and a collective commitment to shaping futures in which GenAI supports – not supplants – human creativity in software development. Taken together, the outcomes of this workshop argue for reframing success in software engineering: from efficiency alone toward creative, humane, and value-aligned systems of work.

2 Table of Contents

Executive Summary

*Victoria Jackson, Rafael Prikladnicki, Daniel Russo, Margaret-Anne Storey, and
André van der Hoek* 23

Introduction 26

Overview of the Workshop 27

Working Sessions 28

Celebrity Interview 30

What, So What, Now What 30

Mad Tea 30

TRIZ: How to Destroy Creativity with GenAI 30

Shift and Share 31

Share Your Passion 31

Short Gallery Walk 31

Wise Crowds 31

STAR – Structure Trust Action Results 31

15% Solutions: Immediate Amplifiers 31

Ecocycle Planning 32

Conversation Café 32

User-Experience Fishbowl 32

1-2-4-All: Generating Research Ideas 32

Dot Voting 33

Why, Who, What, How 33

Open Space 33

Outlining the Dagstuhl Manifesto 33

Wrap Up: Commitment Constellation 34

Manifesto and Next Steps 34

Acknowledgments 35

Participants 36

3 Introduction

The high-level objective of this Perspectives Workshop was to bring together thought leaders from academia and industry to develop a comprehensive research agenda on the intersection of creativity, GenAI, and software development.

Creativity is a bedrock of software development: without it, innovation would stall and the day-to-day problem solving that leads to new features and improvements to existing features would be gravely hampered. Yet, whereas creativity in other disciplines such as arts, engineering, and architecture has been the subject of a vast amount of research, little is known about the nature, importance, and effect of creativity in software development.

The advent of GenAI has thrust the role of creativity in software development to the fore. Some envision a gentle evolution toward GenAI-driven tools that amplify the creative work of developers and make them more effective. Others see a radical near future in which GenAI usurps all rote work, with the sole job left for developers to creatively direct the tools to produce innovative software and features. While a major portion of research, studies, and discussions in the literature and media have concentrated on the impact of GenAI on productivity, this Perspectives Workshop considered that understanding the short-term, medium-term, and long-term impacts of and connections among creativity, GenAI, and software development warrants equal, if not greater, attention.

To discuss, anticipate, and proactively shape the future of creativity augmented by GenAI in software development, this Perspectives Workshop brought together leading academics and practitioner thought leaders to address a range of questions, including: what is the nature of creativity in software development; how might GenAI shape, enhance, and perturb software creativity; what does that mean for the resulting software products; which kinds of novel development tools may be needed; how might organizations and their teams adapt; and what unintended consequences may result?

The workshop had five specific goals:

- Raise awareness of the important role of creativity in software development and the opportunities and pitfalls that may arise with the advent of GenAI.
- Articulate different futures in terms of how GenAI may impact creativity in software development, as bracketed by other relevant factors such as value, productivity, and developer experience and wellbeing.
- Use these futures to develop a comprehensive research agenda that helps shape the future of creativity, GenAI, and software development.
- Ideate new GenAI driven tools that amplify creativity across different software development activities and tasks in positive ways while minimizing potential risks.
- Promote an interdisciplinary community of researchers that cares about, understands, and drives how creativity and GenAI will shape future software development.

Participants in the workshop were specifically selected to bring together expertise in a wide range of backgrounds, including creativity, GenAI, innovation, software analytics, productivity, human-AI teams, design and expertise, software tools, problem solving, and human cognition, among others. To effectively exploit this wide variety of expertise for purposes of future agenda setting, the workshop was run in a highly interactive fashion, without any formal presentations. Instead, participants engaged in many different collaborative activities that were carefully structured and sequenced to build trust, collectively take stock and create common ground, generate many different ideas, articulate possible futures out of those ideas, and refine the most promising directions for future research.

The remainder of this document details how the workshop proceeded, and accompanies the Manifesto that is to be published separately in the first half of 2026. Next, we first provide an overview of the workshop’s approach (section 4), followed by a description of each of the workshop’s working sessions (section 5). The report then concludes with a short preview of the Manifesto’s planned contents (section 6).

4 Overview of the Workshop

The workshop’s overarching goal was to discuss, anticipate, and proactively shape the future of research into the intersection of creativity, Generative AI, and software development. To facilitate discussion, including the identification of future research directions, the workshop was purposefully designed to be interactive and fun, incorporating small-group activities. Many of these activities were drawn from Liberating Structures¹. Liberating Structures are a set of facilitation strategies designed to involve participants in meaningful discussions by breaking conventional patterns, encouraging fresh insights, and collaborative problem-solving. Several of the organizers had prior experience using Liberating Structures in workshops and found them helpful for fostering vibrant discussions for shaping research strategies and directions.

A key output from a Perspectives Workshop such as ours is a Dagstuhl Manifesto. The workshop organizers wished to have an outline of the Manifesto’s contents by the end of the week, which could then be worked on asynchronously by all participants. This required some lightweight structuring of the workshop to move from the many diverse perspectives on creativity among participants to a concrete set of future research directions. The workshop was thus organized with a daily goal, with each day building upon the work and insights from the prior day. The first day’s goal was firmly scene-setting, getting participants to discuss their perspectives on creativity and start thinking about where the future lies. The second day explored the amplification of creative software work through GenAI. The third (short) day began by identifying future research directions, which were subsequently refined on the fourth day, before agreeing on the Manifesto’s contents and starting to contribute to it on the final (short) day.

Table 1 summarizes each of the working sessions that took place during the workshop. As can be seen, there was more structure in the early part of the week, with many sessions to help participants explore the topic. The latter part of the week offered participants an opportunity to provide more details on their preferred research direction.

While the organizers had initially planned the week’s activities, they accepted that the plan might change depending on the outcomes of each day. The organizers met at the end of each day to reflect and decide whether any changes would be required to the upcoming plans. This reflection led to some minor changes, introducing additional Liberating Structures to the planned ones.

On the first day with all the participants, the organizers explained that the workshop was designed to be interactive and stimulating. Liberating Structures were introduced, followed by guidance for participants. Firstly, to create a safe space where participants could openly share their experiences and ideas, the workshop adopted Chatham House rules²: “Participants are free to use the information received, but neither the identity nor the affiliation of the

¹ <https://www.liberatingstructures.com/>

² <https://www.chathamhouse.org/about-us/chatham-house-rule>

■ **Table 1** Summary of the Workshop Activities Ordered by Weekday.

Schedule	Description
Mon. am	Celebrity Interview
Mon. am	What, So What, Now What
Mon. pm	Mad Tea
Mon. pm	TRIZ
Mon. pm	Shift and Share
Tues. am	Share your passion
Tues. am	Short Gallery Walk
Tues. am	Wise Crowds
Tues. am	STAR
Tues. am	15% Solutions: Immediate Amplifiers
Tues. pm	Ecocycle Planning
Tues. pm	Conversation Café
Tues. pm	User-Experience Fishbowl
Wed. am	1-2-4-All
Wed. am	Dot voting
Wed. am	Open Space
Wed. pm	Group trip to Trier
Thurs. am	Why, Who, What, How
Thurs. pm	Open Space
Fri. am	Writing the Dagstuhl Manifesto
Fri. am	Wrap up

speaker(s), nor that of any other participant, may be revealed.” Secondly, the participants were encouraged to engage actively as their insights were crucial to the collective success. We thus asked them throughout the week to contribute by sharing freely, listening intently, and building on others’ ideas. Notably, the whole week’s agenda was not shared with participants to allow the organizers to re-plan as needed.

We hope that by sharing the approach taken to facilitate the collective dialogue necessary to produce a Dagstuhl Manifesto, future workshop organizers can learn from our experience when designing their own Dagstuhl Perspectives Workshop.

The next section shares details of each of the working sections shown in Table 1.

5 Working Sessions

This section describes the various working sessions held during the workshop (as listed in Table 1). The majority of the sessions used techniques drawn from Liberating Structures.

One of the organizers (primarily Daniel Russo) acted as a facilitator to each session. The facilitator introduced the session’s purpose and format, and ensured it stayed within its allotted time. Each session is now described. Accompanying this narrative are images of several artifacts produced during the workshop, as shown in Figure 1.

Importantly, to help grow personal connections and to build positive interactions, each day began with an Impromptu Networking session and ended with a Positive Gossip session. Impromptu Networking asked each participant to pair up and discuss provided questions, such as “What does creativity mean to you?” for several minutes before switching partners to

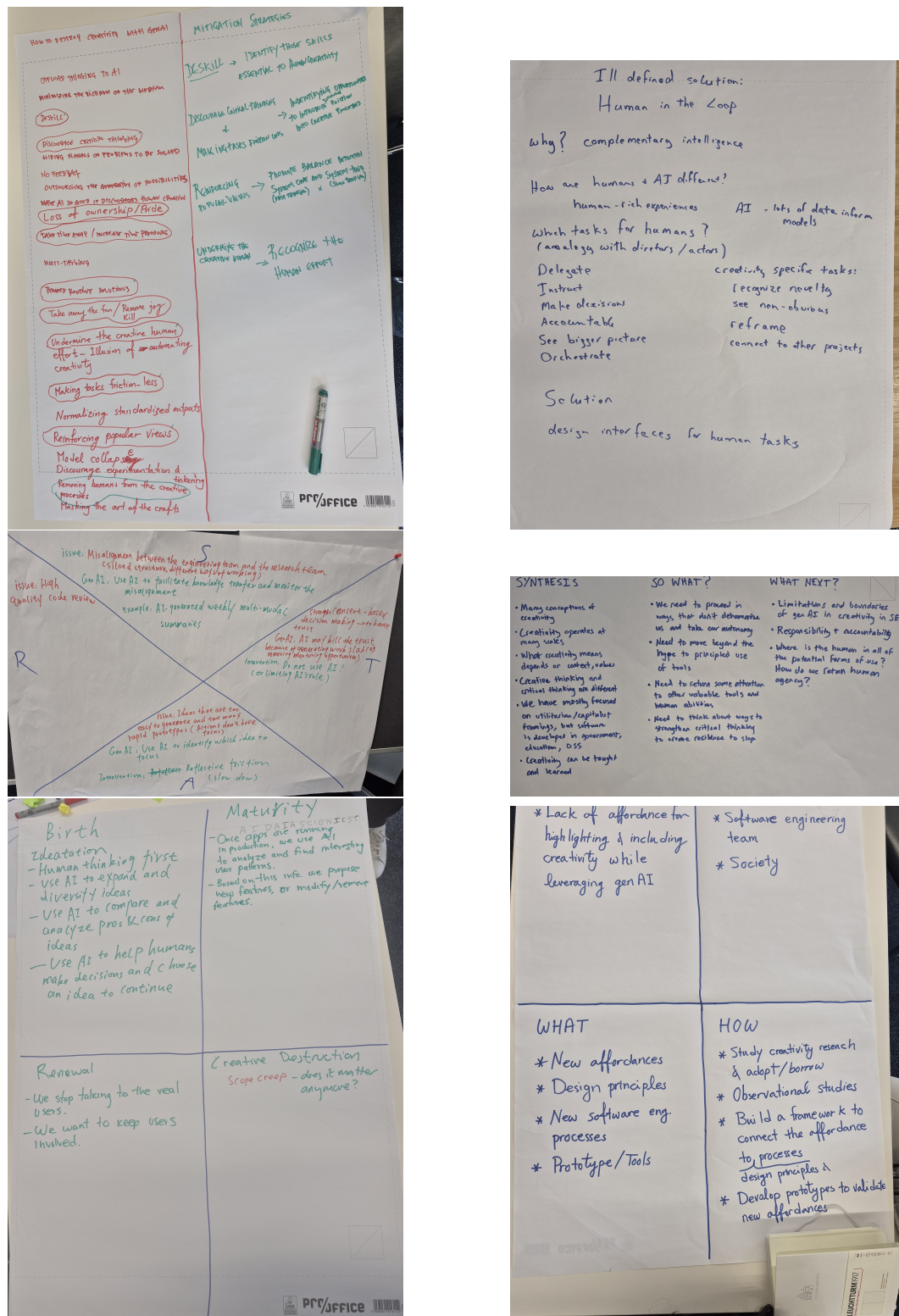


Figure 1 Several Artifacts Produced through the Course of the Workshop.

discuss a different question. “Positive Gossip” asked each participant to first reflect on their own upon the day and identify an idea, moment of curiosity, insight, or a personal quality of someone they appreciated. Participants then formed triads to share “positive gossip” about people outside their triad. The whole group then got together to share emerging themes.

5.1 Celebrity Interview

Creativity is multi-faceted and can involve the creative individual, a creative process, a creative product, or an environment that facilitates creativity. To set the scene for creativity and explore these different facets, the first session was held as a celebrity interview. Three celebrities (participants) were asked to sit on stage and discuss what creativity meant to them. To ensure a broad range of perspectives, two of the participants came from industry (Sarah d’Angelo from Google, Olga Lvova from JetBrains) and one from academia (Victoria Jackson, University of Southampton). The other workshop participants formed five groups and acted as journalists. They listened, took notes, and then each group asked several questions. A lively discussion ensued.

5.2 What, So What, Now What

While the previous session helped to set the scene for what creativity is, this session explored creativity further by asking small groups to answer: “What: Have we surfaced about creativity?”, “So What: Why does it matter right now?” and “Now what: what aspects demands further inquiry?” Each group sat around a table and captured their answers on a poster. Subsequently, the groups came together into a plenary session where each group presented their poster.

5.3 Mad Tea

This session was designed to surface diverse perspectives quickly, using short, provocative prompts in a whole-group circle discussion. Participants formed a circle and took turns responding (briefly) to a set of sentence-starter invitations, including: “Creativity in software engineering means...”, “GenAI helps me think...”, “One concern I have about GenAI in creative work is...”, “Creativity is threatened when...”, and “A tool I wish existed is...”. Contributions were kept concise to maintain momentum and to help the group build shared language around creative practice, risks, and desired tool support.

5.4 TRIZ: How to Destroy Creativity with GenAI

Five groups were formed. Each group was given a poster and asked to create two columns. On the left, they listed all the practices that could be destroyed by AI. They were then asked to circle the ones that they thought would be likely to happen in the next one to two years, and to write mitigation strategies for each circled practice.

5.5 Shift and Share

This session built upon the prior exercise by using the generated posters as input. One person from each table stayed, while the other table members rotated to a different table. The person at the table explained the poster to the newcomers to the table and used sticky notes to capture their feedback. This rotation was repeated three times so that rotating members visited all four tables. This approach enabled people to see perspectives similar to and different from their own. The resulting posters were hung on the room's walls, ready for the next day, as this activity was the final one for the day.

5.6 Share Your Passion

At the start of Tuesday, this short session allowed participants to raise issues relevant to the workshop topic that had not yet been discussed. Each participant was encouraged to write on a sticky note: (1) any topic they cared deeply about, and/or (2) any piece of knowledge or information they felt was relevant to the topic and should be shared. All these sticky notes were captured on a board and referred to throughout the remainder of the week.

5.7 Short Gallery Walk

All participants walked around the room to review the TRIZ posters from the prior day. They were asked to individually pick one ill-defined solution idea from a poster and either keep it as is or build on it a new solution idea that was dear to them.

5.8 Wise Crowds

With the participants split into five groups, each person took a turn discussing their chosen ill-defined solution idea. The other people in the group asked clarifying questions, shared further ideas, and shared their experiences to help progress the ill-defined solution. The resulting ill-defined solutions and feedback were all captured on a poster.

5.9 STAR – Structure Trust Action Results

The goal of this exercise was to reveal how collaboration structure, trust, and action shape creative amplification in organizations. Working in five groups, each group considered each of the STAR dimensions to identify organizational issues and opportunities for GenAI to amplify creativity or where GenAI may erode creativity, and to identify interventions to help strengthen creativity. These were captured on a poster.

5.10 15% Solutions: Immediate Amplifiers

This exercise encouraged participants to turn insights gained during the workshop into small, actionable steps to nurture creative amplification in their sphere of influence. Each participant was asked to consider the prompt, “What can you do tomorrow to let GenAI

amplify creativity in your work? What requires no permission, budget, or policy change?”, and to capture their ideas on a sticky note. Each participant discussed their ideas with another participant to receive and give feedback. The refined ideas were subsequently shared with the wider group by placing them on the “15% wall”.

5.11 Ecocycle Planning

This exercise explored where creativity thrives or stagnates and where GenAI could renew it. To do so, participants reflected on current software development practices and placed them in an ecocycle of birth, maturity, creative destruction, and renewal. Each current practice was briefly described along with its current state and an idea of how GenAI could amplify it. The exercise was first undertaken by a single individual, and then repeated in small groups of five in which the practices were clustered and discussed. Each group then identified two amplification levers to be discussed later.

5.12 Conversation Café

The purpose of this exercise was to encourage reflection on the ethical and human boundaries that sustain meaningful creative amplification. Three rounds of discussion were held in small groups of four to five people. In each round, participants took turns to reflect on the posed question for that round for one minute. The three rounds asked people to consider: (1) Where must human judgment remain central, (2) Where can GenAI genuinely extend creative reach, and (3) What norms of values keep that balance healthy? A fourth round asked for collective synthesis with the shared set of guiding principles for responsible and value-aligned creative amplification captured on a poster.

5.13 User-Experience Fishbowl

This session was an opportunity to ground theory in real-life practices and research. To do so, the industry participants were asked to sit in an inner circle with the academics sitting outside them in an outer circle. The inner circle discussed where GenAI elevated creative quality in their work, when it flattened or distorted creativity instead, and what distinguished the two cases. The people in the outer circle listened to the conversation and, if they had a question or insight, tapped someone in the inner circle on the shoulder to replace them and enter the circle to contribute to the discussion. This created a dynamic of evolving participants and discussion topics.

5.14 1-2-4-All: Generating Research Ideas

Each participant was asked to write one to two bold research questions or project ideas that could meaningfully extend the amplification principles from the prior day. They then shared and discussed these in pairs, and subsequently formed teams of four (by joining another pair) to discuss before placing them on a whiteboard.

All participants then helped to cluster related research questions or project ideas into similar themes.

5.15 Dot Voting

Dot voting was used to prioritize the earlier-identified research ideas. Each participant was given three dots to place on the ideas that resonated most with them. The top five ideas were subsequently refined during the remainder of the workshop to form the basis of the resulting Dagstuhl Manifesto.

5.16 Why, Who, What, How

Each participant chose a research idea to work on and sat with others also interested in that idea. In this activity, which is based upon the research framework proposed by Storey et al.³, the group working on an idea was asked to consider the four points below as they captured their insights on a poster to be hung on the wall and shared more widely for feedback.

- **Why** is this research idea important to study (and which phenomena about the selected idea are especially important to study)?
- **Who** will benefit from this research (e.g., developers, leaders, researchers, tool designers)?
- **What** research contribution will be produced (e.g., descriptive/explanatory/predictive theories, new tools, improved processes)?
- **How** can this research best be performed; that is, what are suitable research strategies that align with the research questions and phenomena to be studied?

5.17 Open Space

In an open space, individuals were free to work on any of the five identified research ideas and to contribute to others. This meant forming small clusters of people interested in a research idea to discuss and refine it further. All insights were captured on a poster to be hung in the shared space.

5.18 Outlining the Dagstuhl Manifesto

Offline discussions over the prior days had revealed a consensus that the Manifesto should be titled “Creativity is the new productivity”. This title was deliberately designed to be provocative and to challenge prevailing views that productivity is the be-all and end-all of software development. The workshop participants agreed that the five identified research directions should serve as the basis for the Manifesto and that they would collectively author it. The final working session spanned the entire morning of the last day and enabled the participants to split into the groups responsible for each section to discuss further and to start capturing information pertinent to their selected idea into a shared Google Doc.

³ Margaret-Anne Storey, Neil A. Ernst, Courtney Williams, Eirini Kalliamvakou *The who, what, how of software engineering research: a socio-technical framework*, Empirical Software Engineering, Volume 25, Issue 5, 4097–4129, 2020, <https://doi.org/10.1007/s10664-020-09858-z>

5.19 Wrap Up: Commitment Constellation

The final session before the group split up for an early lunch and subsequent departure was to visualize ongoing engagement. Each participant wrote one concrete action they will take related to the workshop content and placed it on the floor, near or far from the center depending on its immediacy. Everyone was able to see the constellation and see each other's commitments.

6 Manifesto and Next Steps

A key result from the workshop is a Dagstuhl Manifesto detailing future research directions and places these research directions into a larger societal context. The workshop identified five future research directions to form the basis of the Manifesto (to be published in the first half of 2026). This Manifesto will be collectively authored by all of the workshop participants. The Manifesto is important, but equally important are the connections formed during the workshop. These connections led to various participants planning to collaborate on more detailed research ideas that were identified during the week. Indeed, at the time of this writing, some joint papers have already been authored and submitted. Over time, the group plans to stay connected and to continue to discuss and advance research and practice at the intersection of creativity, GenAI, and software development.

The five research directions to be included in the Manifesto are:

- **Critical perspectives:** Understanding how to develop a science of critical GenAI use in creative software development where biases are understood, intentional architectures are used, multicultural accounts of creativity are incorporated, sustainability is a design parameter, and organization practices and processes guide developers in appropriate use that limits harm.
- **Characterizing creativity:** Understanding what creativity in software engineering looks like for different software development roles and for different aspects of software development, how it plays out in the day-to-day work of software developers as well as at a larger, more organizational scale toward innovation, and how this may differ across individual character traits, team practices, and organizational structures.
- **Evidence of value:** Understanding the value that creativity brings to individuals, teams, and organizations in terms of both process and the resulting product, under what conditions it flourishes or may be hampered, and how it can be effectively promoted and practiced.
- **Design space:** Developing a design space for GenAI-driven tools and processes that promote creativity in development, together with identifying interaction and other design principles that can guide the design of future novel tools and processes that augment and amplify creativity in software engineering.
- **Synthesize research:** How to bring together and package research results and practical knowledge in GenAI, creativity, and software development so to increase the change of widespread impact; this entails effectively synthesizing and communicating emerging results and knowledge in an easily-accessible manner.

7 Acknowledgments

The organizers wish to thank Schloss Dagstuhl, the Leibniz Center for Informatics for hosting and supporting the workshop. Schloss Dagstuhl is a fantastic venue for fostering the lively discussions and collaboration necessary to move research forward. All of the staff was truly awesome in supporting whatever the workshop needed; we are extremely grateful.

Participants

- Aamir Amin
University of Portsmouth, GB
- Souti Chattopadhyay
USC – Los Angeles, US
- Sarah D'Angelo
Google – Mountain View, US
- Peter Dalsgaard
Aarhus University, DK
- Jeanette Falk
Aalborg University
Copenhagen, DK
- Guo Freeman
Clemson University, US
- Jonas Frich
Aarhus University, DK
- Kiev Gama
Federal University of
Pernambuco – Recife, BR
- Cat Hicks
Pluralsight – Draper, US
- Maliheh Izadi
TU Delft, NL
- Victoria Jackson
University of Southampton, GB
- Michael Keeling
Kiavi – San Francisco, US
- Amy Ko
University of Washington –
Seattle, US
- Raymond Li
ServiceNow – Montreal, CA
- Olga Lvova
JetBrains – Yerevan, AM
- Mary Lou Maher
University of North Carolina –
Charlotte, US
- Adriana Meza Soria
MIT-IBM AI Watson Lab –
Cambridge, US
- Gail C. Murphy
University of British Columbia –
Vancouver, CA
- Karim Nakad
Meta Platforms – Bellevue, US
- Marian Petre
The Open University –
Milton Keynes, GB
- Rafael Prikladnicki
Pontificia Universidade Católica
do Rio Grande do Sul –
Porto Alegre, BR
- Daniel Russo
Aalborg University
Copenhagen, DK
- Zohreh Sharafi
Polytechnique Montréal, CA
- Mary Shaw
Carnegie Mellon University –
Pittsburgh, US
- Margaret-Anne Storey
University of Victoria, CA
- André van der Hoek
University of California,
Irvine, US
- Bogdan Vasilescu
Carnegie Mellon University –
Pittsburgh, US



Sound Static Program Analysis in Modern Software Engineering

Pietro Ferrara^{*1}, Liana Hadarean^{*2}, Jorge A. Navas^{*3},
Caterina Urban^{*4}, and Greta Dolcetti^{†5}

1 University of Venice, IT. pietro.ferrara@unive.it

2 Amazon Web Services – Seattle, US. hadarean@amazon.com

3 Certora – Seattle, US. jorge@certora.com

4 Inria & ENS Paris, FR. caterina.urban@inria.fr

5 University of Venice, IT. greta.dolcetti@unive.it

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25421 “Sound Static Program Analysis in Modern Software Engineering”.

Sound Static Program Analysis (SSPA) has historically been effective in proving the absence of runtime errors and security vulnerabilities, notably in safety-critical embedded software. However, it has seen limited adoption in desktop applications until its revival for Web application security (e.g., to prove the absence of SQL injection vulnerabilities). Modern software development, characterized by architectures like microservices, serverless computing, and the increasing use of scripting languages, presents challenges to SSPA due to the integration of multiple languages and complex semantics, while also posing new problems related to soundness, precision, and scalability stemming from the machine learning revolution in code development. Despite these new opportunities for SSPA to offer structured feedback and address serious flaws often overlooked by the shallow analyses currently favored by the industry, there has not been a significant resurgence in its industrial application. This Dagstuhl Seminar aimed to bridge the SSPA and software engineering communities to extend existing theories to these new trends, foster integration with modern practices like DevOps, and discuss the formal methods challenges arising from contemporary software architectures.

Seminar October 12–17, 2025 – <https://www.dagstuhl.de/25421>

2012 ACM Subject Classification Software and its engineering → Automated static analysis;

Software and its engineering → Formal methods; Software and its engineering → Formal software verification; Theory of computation → Program analysis; Software and its engineering → Software verification

Keywords and phrases Abstract interpretation, Formal methods, Software engineering, Software verification, Sound static program analysis

Digital Object Identifier 10.4230/DagRep.15.10.37

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Sound Static Program Analysis in Modern Software Engineering, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 37–74

Editors: Pietro Ferrara, Liana Hadarean, Jorge A. Navas, Caterina Urban, and Greta Dolcetti



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Pietro Ferrara (University of Venice, IT)

Liana Hadarean (Amazon Web Services – Seattle, US)

Jorge A. Navas (Certora – Seattle, US)

Caterina Urban (INRIA & ENS Paris, FR)

License  Creative Commons BY 4.0 International license

© Pietro Ferrara, Liana Hadarean, Jorge A. Navas, and Caterina Urban

Motivation

Sound static program analysis (SSPA) applies formal methods to prove the absence of software defects, including runtime errors and security vulnerabilities. Although its foundational theories – most notably abstract interpretation – were developed nearly half a century ago, their practical adoption has been uneven. SSPA has become indispensable in safety-critical embedded systems, where software failures can have catastrophic consequences, yet it has historically seen limited uptake in desktop and Web applications. This divide has begun to narrow with the growing prominence of Web security threats, such as SQL injection attacks, and the rapid expansion of the Internet of Things (IoT). In today’s software ecosystem – dominated by microservices, serverless architectures, and dynamically typed scripting languages such as Python – SSPA faces a renewed opportunity. Beyond bug prevention, it can offer structured, actionable feedback to a much broader community of developers, including practitioners without expertise in formal verification.

Despite these opportunities, SSPA faces significant challenges regarding the shift toward machine-learning-driven code generation and the complexity of multi-language architectures. Currently, many industrial players rely on shallow, syntactic analyses that leave systems vulnerable to serious flaws. To address this gap, this Dagstuhl Seminar “Sound Static Program Analysis in Modern Software Engineering” (25421) aimed at bridging the SSPA and software engineering scientific communities to adapt formal theories to modern trends. A primary focus was the integration of sound analysis into standard software development lifecycles and DevOps practices.

Summary of Seminar Activities

The Dagstuhl Seminar “Sound Static Program Analysis in Modern Software Engineering” (25421) was attended by 36 researchers, including both senior and junior participants from academia and industry, including graduate students, faculty members, and industry experts. At the beginning of the seminar, each participant briefly introduced themselves and outlined their research interests in a two-minute presentation.

The program interleaved 26 talks (Section 3), two main breakout discussion sessions identifying and discussing open scientific problems (Section 4), and a “speed-dating” session in which participants were divided into *Theory* and *Practice* groups and engaged in a series of short, one-to-one discussions with members of the other group, rotating partners every 10 minutes.

Two invited talks provided perspectives from different points of view at the beginning of the seminar. In particular, Patrick Cousot discussed “Eternal Problems Never or Hardly Solved in Static Analysis by Abstract Interpretation” (Section 3.5), while Davide Taibi tackled “Static Analysis in the Cloud-Native Era” (Section 3.24). Most participants then

illustrated their scientific progress in talks divided into sessions on security, data science programs, machine learning, concurrent software, program verification, heap analysis, and the precision of static analyzers. All together, the talks provided a deep, up-to-date, and some controversial views on the state of the art in SSPA and its application in software engineering practices. This was highly beneficial for sparking discussion of open scientific problems and potential industrial applications. Section 3 reports the abstracts of all the talks.

The breakout discussion sessions were organized into two main steps. During the first breakout session, the participants were randomly split into four distinct groups. Each group then had to identify the most compelling open problems in applying SSA to the software engineering lifecycle. A plenary discussion followed, during which participants identified the four open problems they considered most relevant. The outcome of this discussion is reported in Section 4.1. In a subsequent discussion session, four groups (one for each open problem) were formed, and participants chose which group to join. Each group had a leader and reported the identified problems in a later plenary session. At the end of this discussion, the four most relevant open problems were the standardization of SSA components (Section 4.2), the explainability of SSA results (Section 4.3), the interaction between LLMs and SSA (Section 4.4), and how to push the adoption of SSA at the earlier phases of the software engineering process (Section 4.5).

Last but not least, various social activities took place every evening after dinner. These spanned from organized tournaments to informal board games. Section 5 reports all the results that were tracked during the seminar.

Conclusion

We consider the seminar a success. Altogether, it laid the basis for several tasks and follow-up:

- First of all, each of the four identified open problems represents a fundamental challenge for our community, and each group identified several actionable tasks that we expect the scientific community will target in future work;
- Most of the talks presented either preliminary results not yet published or assessed results with several future perspectives. The lively discussions that took place during the seminar will help the authors to improve the work and continue it in various directions, and other participants will take inspiration for their future work;
- Finally, the friendly atmosphere helped establish new connections through various social activities and informal networking. We expect this will open the door to new collaborations, hopefully leading to novel scientific publications and projects.

2 Table of Contents

Executive Summary

<i>Pietro Ferrara, Liana Hadarean, Jorge A. Navas, and Caterina Urban</i>	38
---	----

Overview of Talks

A Journey through LiSA and its Frontends <i>Vincenzo Arceri</i>	42
Shape Analysis by Abstract Interpretation of Non-Blocking Concurrent Programs <i>Valentin Barbazo</i>	43
Syntactically Convex Model-Based Projection for Linear Rational Arithmetic <i>Anna Becchi and Arie Gurfinkel</i>	43
Distributed Summary Synthesis: A Divide-and-Conquer Approach for Distributed Program Analysis <i>Dirk Beyer</i>	44
Eternal Problems Never or Hardly Solved in Static Analysis by Abstract Interpretation <i>Patrick Cousot</i>	45
Scaling up Roundoff Analysis of Functional Data Structure Programs <i>Eva Darulova</i>	45
Perspectives on the Static Analysis of Synchronous Data-Flow Languages <i>Charles De Haro</i>	46
Path Optimal Symbolic Execution <i>Giovanni Denaro</i>	46
Optional Type Systems: Current Approaches and Ongoing Efforts <i>Werner Dietl</i>	46
Pyra: A High-level Linter for Data Science Software Code Smells <i>Greta Dolcetti</i>	47
RubberDuckBench: A Benchmark for AI Coding Assistants <i>Elizabeth Dinella</i>	48
Securing Software Supply Chains with Security-Enhanced SBOMs <i>Musard Balliu</i>	48
Security Analysis at Scale with the Sigma Engine <i>Isabel Garcia-Contreras</i>	48
Automatic Verification of Replicated Data Types <i>Elisa Gonzalez Boix</i>	49
Directed Program Analysis: from Suspicion to Witnesses <i>Kihong Heo</i>	49
Static Analysis and Verification in The Ciao Playground <i>Manuel Hermenegildo</i>	50
Developing Cost-Effective Combination of Static Analysis Techniques <i>Minseok Jeon</i>	50

Cost of Soundness in Mixed-Precision Tuning <i>Debasmita Lohar</i>	51
Formally Checking the Stability of (Small) Decision Tree Models with Intervals <i>Antoine Miné</i>	51
Try-Mopsa: Relational Static Analysis in Your Pocket <i>Raphaël Monat</i>	52
Static Analysis through Trust Boundaries <i>Naïm Moussaoui Remil</i>	52
Formally Verifying Solana Protocols with the Certora Prover <i>Jorge A. Navas</i>	54
Towards Interactive Abstract Interpretation for Multithreaded Programs <i>Michael Schwarz and Helmut Seidl</i>	55
Static Analysis in the Cloud-Native Era <i>Davide Taibi</i>	55
Tai-e: Sound Static Analysis Framework for Modern Software Engineering <i>Tian Tan</i>	56
From Detection to Quantification: A New Era of Declarative Program Analysis <i>Jingbo Wang</i>	56
Working Groups	
Open Scientific Problems <i>Pietro Ferrara, Liana Hadarean, Jorge Navas, and Caterina Urban</i>	57
Standardization of Static Analysis Components and Interfaces	59
Explainability in Static Analysis	66
Combination of LLMs and Static Analysis	68
Shifting Abstract Interpretation Left: Using Abstraction Interpretation Earlier in the Development Process	69
Social Activities	71
Participants	74

3 Overview of Talks

3.1 A Journey through LiSA and its Frontends

Vincenzo Arceri (*University of Parma, IT*)

License © Creative Commons BY 4.0 International license
© Vincenzo Arceri

Joint work of Vincenzo Arceri, Luca Negrini, Luca Olivieri, Pietro Ferrara, Agostino Cortesi

Main reference Luca Negrini, Pietro Ferrara, Vincenzo Arceri, Agostino Cortesi: “LiSA: A Generic Framework for Multilanguage Static Analysis”, pp. 19–42, Springer, 2023.

URL https://doi.org/10.1007/978-981-19-9601-6_2

LiSA (Library for Static Analysis)[1, 2] is an open-source Java library that provides a comprehensive infrastructure for building static analyzers based on abstract interpretation. We present some of LiSA’s core components, such as the parametric combination of heap and value domains and the CFG nodes-to-symbolic expressions rewriting system. We then present two LiSA frontends, GoLiSA[3] and EVMLiSA[4], which are static analyzers for Go and EVM bytecode, respectively. We discuss the main challenges encountered during the development of both LiSA and its frontends. Finally, we outline the next steps and future directions of the project, such as the integration of string-to-code statement analyses with abstract interpretation in LiSA[5].

References

- 1 Pietro Ferrara, Luca Negrini, Vincenzo Arceri, Agostino Cortesi. Static analysis for dummies: experiencing LiSA. SOAP@PLDI 2021: Proceedings of the 10th ACM SIGPLAN International Workshop on the State Of the Art in Program Analysis, Virtual Event, Canada, 22 June, 2021, pages 1–6, ACM, 2021. 10.1145/3460946.3464316
- 2 Luca Negrini, Pietro Ferrara, Vincenzo Arceri, Agostino Cortesi. LiSA: A Generic Framework for Multilanguage Static Analysis. In: *Challenges of Software Verification*. Ed. by Vincenzo Arceri, Agostino Cortesi, Pietro Ferrara, Martina Oliaro. Intelligent Systems Reference Library, volume 238, pages 19–42, Springer, 2023. 10.1007/978-981-19-9601-6_2
- 3 Luca Olivieri, Luca Negrini, Vincenzo Arceri, Fabio Tagliaferro, Pietro Ferrara, Agostino Cortesi, Fausto Spoto. Information Flow Analysis for Detecting Non-Determinism in Blockchain. 37th European Conference on Object-Oriented Programming, ECOOP 2023, July 17-21, 2023, Seattle, Washington, United States. LIPIcs, volume 263, pages 23:1–23:25, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. 10.4230/LIPIcs.ECOOP.2023.23
- 4 Vincenzo Arceri, Saverio Mattia Merenda, Luca Negrini, Luca Olivieri, Enea Zaffanella. EVMLiSA: Sound Static Control-Flow Graph Construction for EVM Bytecode. *Blockchain: Research and Applications*, pages 100384, 2025. 10.1016/j.bcra.2025.100384
- 5 Vincenzo Arceri, Isabella Mastroeni. A sound abstract interpreter for dynamic code. SAC ’20: The 35th ACM/SIGAPP Symposium on Applied Computing, online event, [Brno, Czech Republic], March 30 – April 3, 2020, pages 1979–1988, ACM, 2020. 10.1145/3341105.3373964

3.2 Shape Analysis by Abstract Interpretation of Non-Blocking Concurrent Programs

Valentin Barbazo (*ENS, PSL University – Paris, FR*)

License  Creative Commons BY 4.0 International license

© Valentin Barbazo

Joint work of Valentin Barbazo, Xavier Rival

While concurrency can achieve significant performance improvements, it also introduces an additional layer of complexity in program design. Synchronization mechanisms are required to prevent data races on shared resources, which could otherwise lead to inconsistent results or data corruption. For instance, pointer updates by one thread may render some memory regions unreachable to others, thereby invalidating their assumptions about shared data. In this talk I present ongoing work toward a shape analysis for concurrent programs based on non-blocking synchronization mechanisms. Our goal is to prove the preservation of structural properties of dynamically allocated data structures shared across multiple threads. Our approach builds on previous work describing how to embed Rely-Guarantee reasoning in the Abstract Interpretation framework [1]. I describe our current progress on the design of an effective thread-modular shape analysis leveraging existing separation-logic-based memory abstract domains to precisely capture and account for thread interactions.

References

- 1 Antoine Miné. “Relational Thread-Modular Static Value Analysis by Abstract Interpretation”. In: *Verification, Model Checking, and Abstract Interpretation*. Ed. by Kenneth L. McMillan and Xavier Rival. Springer Berlin Heidelberg, 2014. DOI: 10.1007/978-3-642-54013-4_3

3.3 Syntactically Convex Model-Based Projection for Linear Rational Arithmetic

Anna Becchi (*USI – Lugano, CH*) and Arie Gurfinkel (*University of Waterloo, CA*)

License  Creative Commons BY 4.0 International license

© Anna Becchi and Arie Gurfinkel

Joint work of Anna Becchi, Arie Gurfinkel, Grigory Fedyukovich, Lev Nachmanson

Quantifier elimination (QE) is a key task in formal verification algorithms. An important feature of QE is the ability to return partial results when interrupted early: a notable example is Model-Based Projection (MBP), a technique used to return an under-approximation of QE. Another desirable feature of QE is that the output is returned in a concise representation. In the theory of Linear Rational Arithmetic (LRA), existing QE methods often fail to preserve syntactic convexity, that is, they return a disjunction even for a conjunctive input, or they return a large non-minimal representation. This can disadvantage the QE client. To address these issues, we introduce the new concept of Bidirectional Model-Based Projection (BMBP). Given a model, BMBP returns both an over- and an under-approximation of QE. We define a new QE algorithm for LRA (BMBP-QE) that (i) is an anytime algorithm capable of providing two complementary partial results – a conjunctive over-approximation and a disjunctive under-approximation – when interrupted early, (ii) returns a minimal conjunction when the input is conjunctive, and (iii) applies to arbitrary LRA formulae. Given the twofold nature of the partial results, BMBP has the potential to enable new strategies in verification

algorithms currently relying on MBP. BMBP also benefits QE clients by returning a minimal representation for conjunctive inputs. We show that our algorithm does not introduce significant overhead compared to QE methods using MBP. In fact, BMBP-QE outperforms SMT-based QE algorithms, offering improvements in both runtime and result size.

3.4 Distributed Summary Synthesis: A Divide-and-Conquer Approach for Distributed Program Analysis

Dirk Beyer (LMU München, DE)

License © Creative Commons BY 4.0 International license

© Dirk Beyer

Joint work of Dirk Beyer, Matthias Kettl, Thomas Lemberger

Main reference Dirk Beyer, Matthias Kettl, Thomas Lemberger: “Decomposing Software Verification using Distributed Summary Synthesis”, Proc. ACM Softw. Eng., Vol. 1(FSE), pp. 1307–1329, 2024.

URL <https://doi.org/10.1145/3660766>

There are many approaches for automated software verification, but they are either imprecise, do not scale well to large systems, or do not sufficiently leverage parallelization. This hinders the integration of software model checking into the development process (continuous integration). We propose an approach to decompose one large verification task into multiple smaller, connected verification tasks, based on blocks in the program control flow. For each block, summaries (block contracts) are computed – based on independent, distributed, continuous refinement by communication between the blocks. The approach iteratively synthesizes preconditions to assume at the block entry (computed from postconditions received from block predecessors, i.e., which program states reach this block) and violation conditions to check at the block exit (computed from violation conditions received from block successors, i.e., which program states lead to a specification violation). This separation of concerns leads to an architecture in which all blocks can be analyzed in parallel, as independent verification problems. Whenever new information (as a postcondition or violation condition) is available from other blocks, the verification can decide to restart with this new information. We formulate our approach as configurable program analysis and implement it for the verification of C programs in the widely used verifier CPAchecker. A large experimental evaluation shows the potential of our new approach: The distribution of the workload to several processing units works well, and there is a significant reduction of the response time when using multiple processing units. There are even cases in which the new approach beats the highly-tuned, existing single-threaded predicate abstraction.

References

- 1 Dirk Beyer, Matthias Kettl, and Thomas Lemberger. Decomposing Software Verification using Distributed Summary Synthesis. Proc. ACM Softw. Eng., 1(FSE), 2024. ACM. doi:10.1145/3660766 <https://doi.org/10.1145/3660766>

3.5 Eternal Problems Never or Hardly Solved in Static Analysis by Abstract Interpretation

Patrick Cousot (New York University, US)

License © Creative Commons BY 4.0 International license
© Patrick Cousot

I tried to remember the history of the discovery of the principles of abstract interpretation and collect problems that remain challenging, some over 50 years, like definition of the semantics of programming languages, efficiency, soundness, scalability, study of the discovery of inductive arguments in inductive proofs, understanding of passage to the limit in infinite domains, study of widening, narrowing, and their duals, absence of tool for the calculational design of abstract semantics, logics, or interpreters by abstraction of a semantics, automation of the design of transformers, exploitation of parallelism in fixpoint computations (asynchronous iterations), necessary non-monotony of widening, separation, modularity, analysis of complex data structures, absence of precise specification of libraries, absence of education in static analysis, incompetence of programmers in the use of static analyzers, explainability, hesitation of hierarchies in front of innovation, short terms goals, ignorance of safety and security, belief that artificial intelligence will solve all problems, proliferation of ad-hoc unprincipled research, deficiencies of reviewers in conferences if not journals, to cite a few.

3.6 Scaling up Roundoff Analysis of Functional Data Structure Programs

Eva Darulova (Uppsala University, SE)

License © Creative Commons BY 4.0 International license
© Eva Darulova

Joint work of Anastasia Isychev, Eva Darulova

Main reference Anastasia Isychev, Eva Darulova: “Scaling up Roundoff Analysis of Functional Data Structure Programs”, in Proc. of the Static Analysis – 30th International Symposium, SAS 2023, Cascais, Portugal, October 22-24, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14284, pp. 371–402, Springer, 2023.

URL https://doi.org/10.1007/978-3-031-44245-2_17

Floating-point arithmetic is counter-intuitive due to inherent rounding errors that potentially occur at every arithmetic operation. A selection of automated tools now exists to ensure correctness of floating-point programs by computing guaranteed bounds on rounding errors at the end of a computation, but these tools effectively consider only straight-line programs over scalar variables. Much of numerical codes, however, use data structures such as lists, arrays or matrices and loops over these. To analyze such programs today, all data structure operations need to be unrolled, manually or by the analyzer, reducing the analysis to straight-line code, ultimately limiting the analyzers’ scalability. We present the first rounding error analysis for numerical programs written over vectors and matrices that leverages the data structure information to speed up the analysis. We facilitate this with our functional domain-specific input language that we design based on a new set of numerical benchmarks that we collect from a variety of domains. Our DSL explicitly carries semantic information that is useful for avoiding duplicate and thus unnecessary analysis steps, as well as enabling abstractions for further speed-ups. Compared to unrolling-based approaches in state-of-the-art tools, our analysis retains adequate accuracy and is able to analyze more benchmarks or is significantly faster, and particularly scales better for larger programs.

3.7 Perspectives on the Static Analysis of Synchronous Data-Flow Languages

Charles De Haro (ENS, PSL University – Paris, FR)

License © Creative Commons BY 4.0 International license
© Charles De Haro

Joint work of Charles De Haro, Marc Pouzet, Xavier Rival

This presentation explores theoretical perspectives on static analyses of synchronous data-flow languages, such as Lustre and SCADE, which model programs as sets of equations describing relationships between input and output streams. We discuss the benefits of analyzing the source program directly, rather than compiling away high-level constructs of these languages – such as automata or clocks – and highlight the advantages of using synchronous observers, which allow safety properties to be specified directly as programs. Additionally, we present scheduling techniques that avoid unnecessary computations when solving equations, while preserving the soundness of the semantics.

3.8 Path Optimal Symbolic Execution

Giovanni Denaro (University of Milano-Bicocca, IT)

License © Creative Commons BY 4.0 International license
© Giovanni Denaro

Joint work of Giovanni Denaro, Pietro Braione, Luca Guglielmo

Symbolic execution is at the core of many techniques for program analysis and test generation. Traditional symbolic execution of programs with numeric inputs enjoys the property of forking as many analysis traces as the number of analyzed program paths, a property that in this talk we refer to as path optimality. On the contrary, current approaches for symbolic execution of heap-manipulating programs fail to satisfy this property, thereby incurring crucial path explosion effects. In the talk, we present Path Optimal Symbolic Execution, a novel symbolic execution algorithm that originally accomplishes path optimality against heap-manipulating programs.

3.9 Optional Type Systems: Current Approaches and Ongoing Efforts

Werner Dietl (University of Waterloo, CA)

License © Creative Commons BY 4.0 International license
© Werner Dietl

The Java type system gives useful guarantees about software, but there are many properties that cannot be expressed. One pervasive property that cannot be expressed is whether a reference can be null or not. The resulting null pointer exceptions are the bane of programmers and have been called the “billion dollar mistake”. They happen even if you think hard about your code and test it thoroughly. There are many causes for null pointer exceptions, including object initialization, missing map keys, confusing contracts, and missing checks. There are many other properties that cannot be expressed in the Java type system that lead to severe runtime errors.

Optional type systems allow developers to improve the quality of their software by encoding additional properties as type systems and enforcing these properties at compile time. We will discuss a type system that prevents null pointer exceptions at compile time and the general framework it builds upon. These optional type systems have found hundreds of bugs in millions of lines of well-tested code. We will discuss alternative tools and interoperability, in particular with Kotlin, and the JSpecify effort to standardize static analysis annotations across the Java ecosystem. Finally, we will briefly discuss whole-program type inference and the interaction of optional type systems with deductive verification.

3.10 Pyra: A High-level Linter for Data Science Software Code Smells

Greta Dolcetti (University of Venice, IT)

License © Creative Commons BY 4.0 International license
© Greta Dolcetti

Joint work of Greta Dolcetti, Vincenzo Arceri, Antonella Mensi, Enea Zaffanella, Caterina Urban, Agostino Cortesi
Main reference Greta Dolcetti, Vincenzo Arceri, Antonella Mensi, Enea Zaffanella, Caterina Urban, Agostino Cortesi: “Introducing Pyra: A High-Level Linter for Data Science Software”, in Proc. of the Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track and Demo Track, pp. 449–453, Springer Nature Switzerland, 2026.
URL https://doi.org/10.1007/978-3-032-06129-4_29

The development of data science software is prone to a variety of mistakes that can easily compromise the validity of results. While existing tools mainly target low-level programming errors, higher-level issues specific to the data science pipeline remain less systematically addressed.

We present Pyra, a high-level linter designed to detect code smells in data science workflows. Pyra relies on static analysis by abstract interpretation, assigning ad hoc abstract datatypes to program variables and checking their consistency when interacting with data science libraries. By adopting a descriptive type system, Pyra captures suspicious patterns without requiring annotations or code modifications, simply by analyzing Python code.

Pyra currently implements 16 checkers across four categories: misleading visualizations, misleading results (e.g., data leakage), challenges for reproducibility (e.g., missing random seeds), and general issues (e.g., unintended consequences of in-place operations).

By bridging programming and domain-specific best practices, Pyra supports more robust, reproducible, and trustworthy data science pipelines.

References

- 1 Dolcetti, G., Arceri, V., Mensi, A., Zaffanella, E., Urban, C., Cortesi, A. (2026). Introducing Pyra: A High-Level Linter for Data Science Software. In: Dutra, I., et al. Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track and Demo Track. ECML PKDD 2025. Lecture Notes in Computer Science(), vol 16022. Springer, Cham. https://doi.org/10.1007/978-3-032-06129-4_29
- 2 Greta Dolcetti, Agostino Cortesi, Caterina Urban, and Enea Zaffanella. 2024. Towards a High Level Linter for Data Science. In Proceedings of the 10th ACM SIGPLAN International Workshop on Numerical and Symbolic Abstract Domains (NSAD '24). Association for Computing Machinery, New York, NY, USA, 18–25. <https://doi.org/10.1145/3689609.3689996>

3.11 RubberDuckBench: A Benchmark for AI Coding Assistants

Elizabeth Dinella (Bryn Mawr College, US)

License © Creative Commons BY 4.0 International license
© Elizabeth Dinella

Joint work of Elizabeth Dinella, Fatma Ayad, Ferida Mohammed, Petros Maniatis, Satish Chandra

Programmers are turning to AI coding assistants to answer questions about their code. Benchmarks are needed to soundly evaluate these systems and understand their performance. To enable such a study, we curate a benchmark of real-world contextualized questions derived from Github pull request comments. Out of this work, we present RubberDuckBench: a multilingual benchmark of questions about code, along with detailed rubrics for evaluating answers. We evaluate a diverse set of 20 LLMs (proprietary & open-source) on answering these questions. We find that even state of the art models fail to give consistent, correct responses across the benchmark. Grok 4 (69.29%), Claude Opus 4 (68.5%), and GPT-5 (67.8%) perform best overall, but do not exhibit pairwise significant superiority over the next 9 best performing models. Most models obtain points through partial credit, with the best performing models only answering at most 2 questions completely correctly across all trials. Furthermore, models often hallucinate with lies in 58.3% of responses on average. Cost analysis reveals no correlation between expense (API pricing or parameter count) and performance. We intend this benchmark to be a target for future research in trustworthy and correct AI coding assistants.

3.12 Securing Software Supply Chains with Security-Enhanced SBOMs

Musard Balliu (KTH Royal Institute of Technology – Stockholm, SE)

License © Creative Commons BY 4.0 International license
© Musard Balliu

Joint work of Eric Cornelissen, Musard Balliu

This talk will discuss an enhancement of software bill of materials (SBOMs) with security capabilities to specify and enforce security policies for applications and their software supply chains with the goal of preventing malware.

References

- 1 Eric Cornelissen and Musard Balliu. *NodeShield: Runtime Enforcement of Security-Enhanced SBOMs for Node.js*. ACM Conference on Computer and Communications Security (CCS 2025), Taipei, Taiwan, 2025

3.13 Security Analysis at Scale with the Sigma Engine

Isabel Garcia-Contreras (Black Duck – Calgary, CA)

License © Creative Commons BY 4.0 International license
© Isabel Garcia-Contreras

We present how we do static analysis in Black Duck using Sigma, a fast static analyzer for web/mobile security. We focus on how to make it easier for customers to analyze their code and then understand and prioritize analysis findings. We demo Sigma on an open-source project.

3.14 Automatic Verification of Replicated Data Types

Elisa Gonzalez Boix (VU – Brussels, BE)

License © Creative Commons BY 4.0 International license
© Elisa Gonzalez Boix

Joint work of Kevin De Porre, Carla Ferreira, Elisa Gonzalez Boix

Main reference Kevin De Porre, Carla Ferreira, Elisa Gonzalez Boix: “VeriFx: Correct Replicated Data Types for the Masses”, in Proc. of the 37th European Conference on Object-Oriented Programming, ECOOP 2023, Seattle, Washington, United States, July 17-21, 2023, LIPIcs, Vol. 263, pp. 9:1–9:45, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

URL <https://doi.org/10.4230/LIPICS.ECOOP.2023.9>

Distributed systems replicate data to improve availability, scalability, and fault tolerance. Ensuring that replicas remain eventually consistent is difficult. Current practices advocate for the use of Replicated Data Types (RDTs) which guarantee convergence out-of-the-box, e.g. CRDTs. However, programming distributed systems using these RDTs is non-trivial due to the lack of appropriate abstractions for replica discovery, update propagation, etc.

In this presentation, we introduce the VeriFx language, a high-level functional object-oriented programming language specially designed to empower developers with automated verification features. VeriFx lets programmers implement RDTs atop functional collections and express correctness properties that are verified automatically. For each proof, VeriFx derives the necessary proof obligations, which are encoded into first-order logic and discharged automatically by leveraging SMT solving. If a property does not hold, VeriFx returns a high-level counterexample. Verified RDTs can be transpiled to mainstream languages, e.g. Scala. VeriFx provides libraries that implement the execution model and define the correctness properties of well-known RDT families. We used those libraries to verify 51 CRDTs and reproduce a study on the correctness of Operational Transformation functions.

3.15 Directed Program Analysis: from Suspicion to Witnesses

Kihong Heo (KAIST – Daejeon, KR)

License © Creative Commons BY 4.0 International license
© Kihong Heo

Joint work of Tae Eun Kim, Jaeseung Choi, Kihong Heo, Sang Kil Cha

Main reference Tae Eun Kim, Jaeseung Choi, Kihong Heo, Sang Kil Cha: “DAFL: Directed Grey-box Fuzzing guided by Data Dependency”, in Proc. of the 32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023, pp. 4931–4948, USENIX Association, 2023.

URL <https://www.usenix.org/conference/usenixsecurity23/presentation/kim-tae-eun>

Static program analysis for software bug detection has steadily advanced over many decades and has become a core technology in modern software development. However, the difficulty of inspecting whether analyzer-reported alarms remains a persistent challenge for many developers. In this talk, we present our journey toward addressing this issue and realizing a “fully automated bug detection tool” including buggy input generation. The key lies in “directed input generation”. This technique aims to automatically generate inputs that reach suspicious program points (e.g., static analysis alarms, recently changed code, etc). We will discuss how we combine static analysis with fuzzing and program synthesis to efficiently generate bug-triggering inputs for such points. Also, we share the remaining challenges to improve the current state-of-the-art.

3.16 Static Analysis and Verification in The Ciao Playground

Manuel Hermenegildo (IMDEA Software Institute – Madrid, ES)

License © Creative Commons BY 4.0 International license
© Manuel Hermenegildo

Joint work of Manuel Hermenegildo, José Francisco Morales, Pedro López, Daniel Jurjo, Marco Ciccalè, Daniela Ferreiro, Marco Perez, Louis Rustenholz

URL <http://play.ciao-lang.org>

We presented and demoed the Static Analysis and Verification functionality currently supported within the Ciao playground (<https://play.ciao-lang.org>). The playground is a browser-based program development environment which integrates the CiaoPP abstract interpreter and verifier as an embedded tool. In contrast with the more traditional ways of using CiaoPP (within Visual Studio Code, Emacs, etc.), the playground does not require any installation, while retaining almost all the capabilities of the full, native version of the systems involved. This makes it a very convenient tool for experimentation and teaching. The whole system runs locally on the student's browser, compiled to web assembly code, and does not need any server infrastructure. This has advantages from the point of view of scalability, low maintenance cost, security, privacy, etc. We presented the playground in action, in different analysis and verification including interactive, on-the-fly, assertion checking as the program is developed, covering properties such as types/shapes, sharing/aliasing, numeric properties, determinism, data sizes, computational cost, etc. We also showed the different configuration options available and the different kinds of information that can be obtained from the analyzer. We demonstrated as well the abstract debugger, that allows advancing step by step in the source program editor while watching the evolution of the abstract values as the fixpoints are computed. Finally, we presented how documents with embedded instances of the analyzer can also be developed within the playground, which allows easily developing interactive tutorials for static analysis tasks, from analyzing and verifying programs to developing new abstract domains.

3.17 Developing Cost-Effective Combination of Static Analysis Techniques

Minseok Jeon (DGIST Institute of Science & Technology – Daegu, KR)

License © Creative Commons BY 4.0 International license
© Minseok Jeon

In this talk, I will introduce an issue in the current trend of developing static analysis techniques and present an efficient approach to address it. To develop cost-effective static analyzers, various static analysis techniques have been proposed. For optimal performance, these techniques should be combined to leverage the benefits of each approach. In the literature, however, the techniques have largely been developed independently, without considering their potential combinations. We discovered that combining independently developed static analysis techniques often results in suboptimal performance, even when each technique is individually optimal. However, developing combinations of techniques while accounting for all interactions between them is a huge burden. To address this issue, I introduce an approach that safely and effectively reduces the burden of developing such combinations.

3.18 Cost of Soundness in Mixed-Precision Tuning

Debasmita Lohar (KIT – Karlsruher Institut für Technologie, DE)

License © Creative Commons BY 4.0 International license
© Debasmita Lohar

Joint work of Debasmita Lohar, Anastasia Isychev

Main reference Anastasia Isychev, Debasmita Lohar: “Cost of Soundness in Mixed-Precision Tuning”, Proc. ACM Program. Lang., Vol. 9(OOPSLA2), Association for Computing Machinery, 2025.

URL <https://doi.org/10.1145/3763137>

Mixed-precision tuning is a key optimization technique for numerical programs, where some variables and operations are assigned lower precisions to improve overall performance. It is particularly relevant in applications with frequent numerical computation on resource-constrained hardware, such as embedded systems, scientific computing, and machine learning. The main challenge lies in the trade-off: while reduced precision improves efficiency, it introduces rounding errors that can compromise accuracy.

In this work, we present the first comprehensive evaluation of state-of-the-art tools for mixed-precision tuning. These tools typically fall into two categories: sound static analyses, which offer formal guarantees but are often considered overly conservative, and dynamic methods, which explore a broader optimization space but can violate error bounds, sometimes by several orders of magnitude. Through an extensive comparison on the FPBench benchmark suite, we quantify the trade-offs between performance and soundness. Our results show that sound tools, when enhanced with techniques such as regime inference, can match or even outperform dynamic ones, while still preserving correctness guarantees.

Yet, extending sound guarantees to larger, real-world programs remains a challenge. A promising direction could be to combine static analyses with heuristic guidance – potentially informed by large language models that can help navigate the search space – to improve scalability and enable integration into real-world software engineering workflows.

3.19 Formally Checking the Stability of (Small) Decision Tree Models with Intervals

Antoine Miné (Sorbonne University – Paris, FR)

License © Creative Commons BY 4.0 International license
© Antoine Miné

Joint work of Waly Fall, Antoine Miné

Decision Trees are a popular kind of predictive models used in machine learning. Previous work in explainable AI by Hurault and Marques-Silva proposed logic-based criteria and algorithms to provide local explanations for model decisions (so-called abductive and contrastive explanations), but efficient explanation synthesis requires models to be stable, a global property difficult to check. In this work, we develop an interval-based method able to formally prove (small-scale) decision tree models to be stable. When they are not stable, we are able to generate counter-examples, compute a quantitative measure of stability, and even provide fixes to make the model stable. We also consider monotony, a more restrictive property than stability, and prove formally monotony as well as infer, when possible, all orders that make the model monotone. We present preliminary experimental results on a few small (4-7 features, 50KB-2MB model size) models trained using Gradient Boosting with XGBoost and LightGBM, from data from OpenML, Kaggle, and UC Irvine. Our preliminary results indicate that, even though the training algorithms are advertised as

computing monotone models, the output is not always monotone, and not even stable. In some cases, almost-stable models can be made stable while keeping (or improving) their size and precision. We also show that one-hot encoding is useful to ensure monotony and stability of trained models. This is joint work with Waly Fall.

3.20 Try-Mopsa: Relational Static Analysis in Your Pocket

Raphaël Monat (INRIA Lille, FR)

License © Creative Commons BY 4.0 International license
© Raphaël Monat

Main reference Raphaël Monat: “Try-Mopsa: Relational Static Analysis in Your Pocket”, CoRR, Vol. abs/2509.13128, 2025.

URL <https://doi.org/10.48550/ARXIV.2509.13128>

Static analyzers are complex pieces of software with large dependencies. They can be difficult to install, which hinders adoption and creates barriers for students learning static analysis. This work introduces Try-Mopsa: a scaled-down version of the Mopsa static analysis platform, compiled into JavaScript to run purely as a client-side application in web browsers. Try-Mopsa provides a responsive interface that works on both desktop and mobile devices. Try-Mopsa features all the core components of Mopsa. In particular, it supports relational numerical domains. We present the interface, changes and adaptations required to have a pure JavaScript version of Mopsa. We envision Try-Mopsa as a convenient platform for onboarding or teaching purposes.

3.21 Static Analysis through Trust Boundaries

Naïm Moussaoui Remil (ENS, PSL University – Paris, FR)

License © Creative Commons BY 4.0 International license
© Naïm Moussaoui Remil

Joint work of Naïm Moussaoui Remil, Caterina Urban

Background

Classical static analysis methods used to treat variables behavior uniformly. Yet, in practice, this can be unnecessarily restrictive. Indeed, not all programs variables play the same role. Some variables are *untrusted*: their value may be externally-controlled, for instance by user-input, an operating system scheduler, or a third-party library. By contrast, other variables are *trusted*: their values stem from non-deterministic choices, internal computations, or library calls under the control of the programmer. Recent works have studied safety properties while distinguishing trusted variables untrusted variables. Girol et al. [1] proposed Robust Reachability which refine the classical notion of reachability. Robust reachability holds if the untrusted variable can make a bug reachable whatever the values of the trusted variables. Also, Parolini et al. [2] define a static analysis for *safety* Non-Exploitability proving that the untrusted variables cannot trigger or silence a safety bug. We go beyond safety properties, and contribute with static analysis for (generalization of) liveness and CTL properties withing this setting.

Presentation Proposal

First, we will present an abstract-interpretation based static analysis that proves that for every possible (sequence of) untrusted input(s), there exists at least one terminating execution. We call this property Termination Resilience, it generalizes termination on program with untrusted and trusted inputs. We also propose a static analysis Robust Non-Termination which is defined as the negation of Termination Resilience i.e. there exists a choice for the untrusted inputs such that the program always diverge. Next, we will present abstract-interpretation based static analysis for inferring minimal sets of inputs variables, along with a sufficient preconditions, ensuring a CTL (computational tree logic) property. This static analysis can be interpreted in multiple way – for example, as a partition of the inputs variables in untrusted and trusted variables that guarante the robustness or the resilience of a CTL property.

Termination Resilience

When the control of some untrusted variables alone is sufficient to enforce divergence, we say that the program is vulnerable to Robust Non-Termination, which represents a serious concern in practice. Non-termination that may be avoided through trusted choices may be acceptable, as it cannot be reliably exploited by an adversary; it reflects internal system behavior rather than external influence. From a security perspective, this distinction is crucial: divergence caused by untrusted inputs constitutes a potential denial-of-service vulnerability, whereas divergence caused by trusted variables does not. From a software engineering standpoint, the distinction is equally important as it hints at a principled way to triage non-termination alarms, helping developers prioritize the more critical cases.

This distinction motivates the property of Termination Resilience; for every possible (sequence of) untrusted inputs, there exists at least one terminating execution. We have derived a sound static analysis for both Termination Resilience and Robust-Non Termination by building upon Urban and Miné’s decision tree abstract domain [4, 5], enhancing it with novel and non-trivial transformer operators to effectively manage this mixed settings. When unable to prove that a program always satisfies Termination Resilience, our static analysis automatically infers *sufficient preconditions* that ensure Termination Resilience.

Inference of Minimal Sets of Variables Ensuring a CTL Property

We have extended an abstract-interpretation-based static analysis for CTL properties (using the decision tree abstract domain) with an abstraction refinement process. This abstract refinement process infers minimal sets of variables, along with sufficient preconditions, to ensure a CTL property. One interpretation of this analysis is the inference of partitions of the input variables into untrusted and trusted sets for a given CTL property. For instance, if the CTL property defines the reachability of a bug, our analysis infers minimal sets of input variables with sufficient preconditions that ensure the bug is reachable. Hence, our analysis automatically infers sets of untrusted input variables with sufficient preconditions that ensure Robust Reachability holds.

This work has been published in the 25th Conference on Logic for Programming, Artificial Intelligence and Reasoning [3].

References

- 1 Guillaume Girol, Benjamin Farinier, Sébastien Bardin. Not All Bugs Are Created Equal, But Robust Reachability Can Tell the Difference. In: *Computer Aided Verification – 33rd International Conference, CAV 2021, Virtual Event, July 20-23, 2021, Proceedings, Part I*. Ed. by Alexandra Silva and K. Rustan M. Leino. Lecture Notes in Computer Science, volume 12759, pages 669–693, Springer, 2021. 10.1007/978-3-030-81685-8_32
- 2 Francesco Parolini, Antoine Miné. Sound Abstract Nonexploitability Analysis. In: *Verification, Model Checking, and Abstract Interpretation – 25th International Conference, VMCAI 2024, London, United Kingdom, January 15-16, 2024, Proceedings, Part II*. Ed. by Rayna Dimitrova, Ori Lahav, and Sebastian Wolff. Lecture Notes in Computer Science, volume 14500, pages 314–337, Springer, 2024. 10.1007/978-3-031-50521-8_15
- 3 Naïm Moussaoui Remil, Caterina Urban, Antoine Miné. Automatic detection of vulnerable variables for CTL properties of programs. In: *25th Conference on Logic for Programming, Artificial Intelligence and Reasoning*. Volume 100, pages 116–126, EasyChair, 2024.
- 4 Caterina Urban, Antoine Miné. A decision tree abstract domain for proving conditional termination. *International Static Analysis Symposium*, pages 302–318, Springer, 2014.
- 5 Caterina Urban, Samuel Ueltschi, Peter Müller. Abstract Interpretation of CTL Properties. *SAS 2018*, pages 402–422, 2018.

3.22 Formally Verifying Solana Protocols with the Certora Prover

Jorge A. Navas (Certora – Seattle, US)

License © Creative Commons BY 4.0 International license

© Jorge A. Navas

Joint work of Jorge A. Navas, Arie Gurfinkel

The Solana Certora Prover (SCP) is a formal verification tool designed to check the correctness of Solana smart contracts. SCP takes as input both a Solana program and its specification, both written in Rust, and determines whether the implementation satisfies its specification.

In this seminar, I will demonstrate how SCP can be applied to prove fundamental correctness properties of a tokenized vault protocol, including solvency, absence of shares dilution, and fee assessment. I will also discuss the main technical challenges involved in generating efficient and sound SMT-based verification conditions, and how Abstract Interpretation techniques are used to address these challenges.

3.23 Towards Interactive Abstract Interpretation for Multithreaded Programs

Michael Schwarz (National University of Singapore, SG) and Helmut Seidl (TU München – Garching, DE)

License © Creative Commons BY 4.0 International license

© Michael Schwarz and Helmut Seidl

Joint work of Julian Erhard, Simmo Saan, Sarah Tilscher, Michael Schwarz, Karoliine Holter, Vesal Vojdani, Helmut Seidl

Main reference Julian Erhard, Simmo Saan, Sarah Tilscher, Michael Schwarz, Karoliine Holter, Vesal Vojdani, Helmut Seidl: “Interactive abstract interpretation: reanalyzing multithreaded C programs for cheap”, *Int. J. Softw. Tools Technol. Transf.*, Vol. 26(6), pp. 647–667, 2024.

URL <https://doi.org/10.1007/S10009-024-00768-9>

To put sound program analysis at the fingertips of working class developers, we propose a framework for interactive abstract interpretation of multithreaded C code leveraging mixed-flow sensitivity. Abstract interpretation provides sound analysis results, but can be quite costly in general. To achieve quick response times, we incrementalize the analysis infrastructure, including post-processing, without necessitating any modifications to the analysis specifications themselves. The integration of update rules enables precise incremental analysis of intricate program properties – including concurrency deficiencies such as data-races and deadlocks. A prototype of the framework has been implemented in the static analyzer *Goblint*, and integrated into IDEs leveraging *MagpieBridge*. We evaluate our implementation w.r.t. the yard sticks of response time and from-scratch-consistency. We provide examples of program development highlighting the usability of our approach. Lastly, we discuss promising new directions along the lines of on-demand precision increases and incremental result refinement as an interaction between the analyzer and a developer, making the analysis not just incremental but *interactive*.

3.24 Static Analysis in the Cloud-Native Era

Davide Taibi (University of Southern Denmark – Odense, DK)

License © Creative Commons BY 4.0 International license

© Davide Taibi

Modern cloud-native systems are built as polyglot: multiple languages, frameworks, containers, and deployment descriptors. This makes their actual architecture hard to see and even harder to verify against intended blueprints.

This talk presents a static, multi-artifact analysis approach that reconstructs a system’s architecture by mining source code, build files, Dockerfiles, Helm/Kubernetes manifests, API contracts, and developers’ communication. In particular, the reconstruction enables to detect recurring cloud-native antipatterns such as shared databases or cyclic service dependencies and to have an overview on the system evolution

3.25 Tai-e: Sound Static Analysis Framework for Modern Software Engineering

Tian Tan (Nanjing University, CN)

License © Creative Commons BY 4.0 International license

© Tian Tan

Main reference Tian Tan, Yue Li: “Tai-e: A Developer-Friendly Static Analysis Framework for Java by Harnessing the Good Designs of Classics”, in Proc. of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2023, Seattle, WA, USA, July 17-21, 2023, pp. 1093–1105, ACM, 2023.

URL <https://doi.org/10.1145/3597926.3598120>

Main reference Yufei Liang, Teng Zhang, Ganlin Li, Tian Tan, Chang Xu, Chun Cao, Xiaoxing Ma, Yue Li: “Pointer Analysis for Database-Backed Applications”, Proc. ACM Program. Lang., Vol. 9(PLDI), pp. 1417–1441, 2025.

URL <https://doi.org/10.1145/3729307>

Main reference Menglong Chen, Tian Tan, Minxue Pan, Yue Li: “PacDroid: A Pointer-Analysis-Centric Framework for Security Vulnerabilities in Android Apps”, in Proc. of the 47th IEEE/ACM International Conference on Software Engineering, ICSE 2025, Ottawa, ON, Canada, April 26 – May 6, 2025, pp. 2803–2815, IEEE, 2025.

URL <https://doi.org/10.1109/ICSE55347.2025.00208>

Main reference Teng Zhang, Yufei Liang, Ganlin Li, Tian Tan, Chang Xu, Yue Li: “Bridge the Islands: Pointer Analysis for Microservice Systems”, Proc. ACM Softw. Eng., Vol. 2(ISSTA), pp. 504–526, 2025.

URL <https://doi.org/10.1145/3728896>

Static analysis framework plays a crucial role in determining the capabilities of static analysis to adapt and integrate with modern software engineering practices. Recognizing the limitations of existing frameworks, we have developed Tai-e, a new static analysis framework for Java, aimed at meeting the demands of contemporary software development. Our study shows that Tai-e is easier to learn and use, offering a more extensible and efficient pointer analysis system than other frameworks. This provides a solid foundation for the development and integration of new analysis techniques.

A distinctive feature of modern software engineering today is the reliance on various frameworks, such as microservice, database, and mobile frameworks, rather than building software from scratch. While these frameworks greatly enhance productivity, their complexity poses challenges to the soundness of static analysis. To address this, we leveraged Tai-e’s extensible pointer analysis system to develop a series of innovative techniques and tools for effectively analyzing these framework-based applications. Our evaluation demonstrates that our techniques achieve significantly better soundness compared to state-of-the-art methods for these types of applications.

In this talk, I will introduce Tai-e and our techniques for analyzing modern framework-based software.

3.26 From Detection to Quantification: A New Era of Declarative Program Analysis

Jingbo Wang (Purdue University – West Lafayette, US)

License © Creative Commons BY 4.0 International license

© Jingbo Wang

Declarative program analysis has long been a powerful tool for detecting data races, tracking information flows, and uncovering security vulnerabilities such as side-channel leaks. However, designing analysis rules that are both accurate and efficient remains a significant challenge, even for domain experts. To improve scalability, many prior works favor efficiency by relaxing completeness, leading to rules that are sound but incomplete. While this trade-off improves performance, it also produces a flood of false alarms, overwhelming developers and reducing the practical usefulness of the analysis.

In this talk, I will present our approach to bridging qualitative and quantitative program analysis. Rather than simply reporting whether a potential issue exists, our framework estimates the probability that a reported issue – such as an information leak – corresponds to a real vulnerability. This probabilistic perspective enables the analysis to prioritize high-risk findings and significantly reduce false alarms. Our solution is built on probabilistic logic programming, extending the widely used Soufflé Datalog engine. I will demonstrate how this framework achieves efficiency, accuracy, and incremental adaptability, making it well-suited for evolving codebases and complex analysis domains.

4 Working Groups

4.1 Open Scientific Problems

Pietro Ferrara (University of Venice, IT)

License © Creative Commons BY 4.0 International license
© Pietro Ferrara, Liana Hadarean, Jorge Navas, and Caterina Urban

On the first day, participants were randomly split into four groups. Each group identified a leader and was supervised by one of the organizers. Each group discussed and then identified the three most compelling open scientific problems in applying static program analysis to modern software.

4.1.1 Group 1

Leader: Helmut Seidl, Organizer: Pietro Ferrara, Participants: Vincenzo Arceri, Musard Balliu, Werner Dietl, Arie Gurfinkel, Debasmitha Lohar, Naïm Moussaoui Remil, Tian Tan

The discussion was somewhat controversial. First of all, formalizing a sound semantics for a real-world programming language is highly complex and time-consuming, given the wide variety of constructs, primitives, and technologies in each programming language. On the other hand, the best approach to explaining formal methods is to apply them to minimal languages, highlighting the main features tackled by the formalization. So far, it seems there is no solution to reconcile these two opposing needs. In addition, modern software comprises large libraries, and modular reasoning is needed to analyze them. In such a context, several relevant open problems need to be solved: how can we model these libraries? How can we verify their implementation meets their specification? Can we automatically derive the specification?

Open problems

- 1.1 Language workbench
- 1.2 Tiny languages
- 1.3 Modularity

4.1.2 Group 2

Leader: Michael Schwarz, Organizer: Caterina Urban, Participants: Roberto Bagnara, Patrick Cousot, Isabel Garcia-Contreras, Elisa Gonzalez Boix, Kihong Heo, Antoine Miné, Michael Schwarz

The first open problem concerns systems that comprise different languages and paradigms, and in particular, how these systems can be effectively statically analyzed without duplicating the effort of the scientific community. Other communities, particularly those focused on SMTs, have already successfully tackled similar problems and might inspire our community.

Another problem concerns the explainability and reproducibility of warnings. In particular, a warning by itself is not enough in production, since the software developer or project manager might not be able to understand the cause of the warning.

Finally, the last aspect concerns the application of static analysis at the early (aka, left) phases of the development lifecycle (e.g., during development) rather than at the end (e.g., before deployment). This requires incremental and/or interactive analysis and would help developers produce more robust code.

Open problems

- 2.1 Multilanguage and multiparadigm systems
- 2.2 Explainability of warnings
- 2.3 Shifting static analysis left

4.1.3 Group 3

Leader: Greta Dolcetti, Organizer: Jorge A. Navas, Participants: Valentin Barbazov, Eva Darulova, Charles De Haro, David Delmas, Minseok Jeon, Tim King, Valentina Lenarduzzi

The discussion began with the adoption of large language models (LLMs) in static analyzers. This is a recent, emerging scientific trend, but it is not yet clear how these can be adopted to help configure and use static analyzers. To explain their warnings? In combination or in replacement of static analysis? In addition, several properties (e.g., hallucinations) of LLMs have not yet been formalized, and their sustainability should be considered. A second open problem was identified in the lack of non-linear abstract domains. While there are very few examples, it seems that specific domains (such as avionics and machine learning algorithms) would need much more generic and expressive numerical domains in this context. Finally, two more open problems were discussed, one concerning shared data structures in concurrent or distributed systems and how components interact through them, and another one regarding differential analysis and portability.

Open problems

- 3.1 Integration of LLM into static program analysis
- 3.2 Non-linear abstract domains
- 3.3 Shared data structures in concurrent/distributed systems
- 3.4 Differential analysis and portability

4.1.4 Group 4

Leader: Raphaël Monat, Organizer: Liana Hadarean, Participants: Anna Becchi, Dirk Beyer, Giovanni Denaro, Elizabeth Dinella, Guido Salvaneschi, Jingbo Wang

Three main topics were identified from the beginning. Building on the idea of witnesses adopted by competitions for software verification tools, the group identified the need for the certification of programs through witnesses that are cheaper than complete analysis but still prove that the programs respect given properties. A second topic concerned how different analysis engines (potentially implemented with different programming languages and technologies) can cooperate, and how this can be generalized. Last but not least, how to communicate the results of an analysis and its justifications to non-expert users is still obscure and non-standard, as well as making the analysis assumptions clear and explicit to make the analysis more transparent.

Open problems

- 4.1 Certification of programs for sound AI use.
- 4.2 Cooperative verification engines
- 4.3 User interface and communication with non-expert users

4.1.5 Plenary Discussion

All the open problems identified by the various groups were grouped into homogeneous topics. Then each topic was voted on by the participants, and the four most voted on were chosen. Each participant was free to vote on many topics. The final results were as follows:

- A Standardized multilanguage static analysis (open problems: 1.3, 2.1, 4.2), 18 votes.
- B Explainability (open problems: 4.1, 4.3, 2.2), 20 votes.
- C LLMs and static program analysis (open problems: 3.1), 15 votes.
- D Real-world huge vs. academic tiny languages (open problems: 1.1, 1.2), 8 votes.
- E Shifting static analysis left (open problems: 2.3), 13 votes.
- F Non-linear abstract domains (open problems: 3.2), 7 votes.
- G Shared data structures in concurrent/distributed systems (open problems: 3.3), 9 votes.
- H Differential analysis and portability (open problems: 3.4), 4 votes.

Therefore, topics A, B, C, and E were chosen. Similar to the first selection phase, a leader and an organizer were chosen for each topic, but participants were free to join the topic they preferred. The outcomes of each discussion group are presented in Sections 4.2, 4.3, 4.4, and 4.5.

4.2 Standardization of Static Analysis Components and Interfaces

Discussion Leader: Raphaël Monat, Inria & University of Lille

We discussed how to standardize static analysis components and their interfaces, following the standardization successes in neighboring communities such as SMT-LIB [5] in the last decade. These challenges have been identified for more than 20 years by [6, 7] in their “Verifying compiler challenge” and “Verified software initiative”.

This standardization can be achieved at different granularities, e.g at the whole analyzer level, or for given abstract domains (components of the analyzer).

Standardizing inputs and outputs at the analyzer level would be a good first step. It would allow to share and compare results, and even interoperate between analyses. Standardizing abstract domains would be more difficult, as their interfaces currently vary a lot depending on the kind of abstraction. From a technical point of view, such an interface should be supported by the wide variety of implementation languages (C/C++, Java, OCaml, Prolog) used by the community. Thanks to standardization, maintenance efforts of components could be shared by the whole community (rather than each group having to re-implement and maintain classical components).

We identified several efforts paving the way towards a kind of standardization:

- the Apron [12] library for relational abstract domains offers a de-facto interface supported by a wide variety of relational libraries, including PPL [13] Elina [14] and PPLite [16]. Some additional strong points underlined by participants where the fact that this interface is available through multiple language bindings (C/Java/OCaml), and that it is actively maintained.
- external fixpoint solvers such as the top-down solver [8, 19] could also be used by multiple tools. They are currently used in Goblint [33, 34] and Salto [32].
- correctness and violation witness [11] used in the software verification competition. A new format version 2.0 [20] was introduced in 2024. They allow the encoding of invariants and contracts – expressed through side-effect-free C expressions, although the format could be extended to e.g. ACSL [28] – or error paths. The format was extended to non-termination [27] in 2025. There is ongoing work to extend the format to handle more properties: memory and termination, as well as concurrency [15].
- the Static Analysis Results Interchange Format (SARIF) [29] format is an industrial format integrated within CI processes. It is extremely general and does not specify how different analyses of a same language could communicate.
- recent intermediate verification-oriented languages have been proposed, by [30] in the field of software verification, [10] in the field of deductive verification, and by [17] in model checking.
- JavaSMT [9] and PySMT [18] offer features of SMT solvers for software development in Java and Python, as an alternative to the SMT-LIB format.
- modular analyses platforms:
 - Crab [21] is a modular C++ library exposing reusable components such as abstract domains (with integration for PPL [13], Apron [12], and Elina [14]), fixpoint solvers, inter-procedural algorithms, and a language-agnostic Intermediate Representation (IR) enabling effective static analysis across diverse contexts. First, it serves as the foundation for stand-alone static analyzers or verifiers such as Clam [31], a LLVM-based abstract interpreter that translates LLVM bitcode to Crab IR, and Prevail [23], a verifier for eBPF programs. Second, it is used as an evaluation framework by other researchers where new abstract domains or algorithms have been implemented (see e.g., [14, 24, 25, 26]). Third, it integrates within broader verification infrastructures: Clam is part of the verification framework SeaHorn [22], enabling a combination of Abstract Interpretation with Model-Checking techniques. This versatility facilitates the standardization of static analyses by offering common building blocks that can be reused across tools and research projects.
 - Mopsa [36, 35] is a modular and open static analysis platform written in OCaml. Its goal is to encourage the research and education in abstract interpretation by providing a fully-featured and extensible open-source platform and usable analyses built with it. It supports the analysis of multiple languages (C [44, 42], Python [45], OCaml [47]), and

even multi-language analyses [46]. These analyses share the usage of some core abstract domains defining a semantic kernel. Mopsa has been used to develop new kind of analyses including non-exploitability of bugs [39], sufficient precondition inference [40], as well as patch and endianness portability analyses in industrial settings [38, 37]. In terms of usability, Mopsa provides transparency by reporting the proofs it attempted, supports abstract debugging [41] and is available in lightweight web-based version for demonstration purposes [43]. Mopsa has recently been used as a building block by researchers within the software engineering community [50, 49, 48].

- CiaoPP [1, 2, 3] is both an abstract interpretation-based pre-processor and a library of components for building abstract interpretation-based tools. As a tool CiaoPP can perform interactive program analysis, verification, certification (abstraction-carrying code), and optimization (abstract partial evaluation, slicing, program parallelization). The analysis process can be followed interactively on the program source. Different source languages can be supported by translation to a Horn clause-based IR. The tool is highly configurable in order to make use of all the different library components. Regarding the library, it includes a good number of abstract domains for shapes/types, variable (pointer) sharing and other variable instantiation properties, numerical properties, non-failure, determinacy, bounds on computational cost, bounds on data sizes, etc. There are also facilities for defining and combining domains and interfaces with tools like PPL. The library also includes several different analysis fixpoint algorithms (including incremental and modular), abstract partial evaluators, parallelization modules, etc. An almost complete version of CiaoPP can be run interactively on the browser [4].

References

- 1 M. V. Hermenegildo, G. Puebla, F. Bueno, and P. Lopez Garcia. Integrated Program Debugging, Verification, and Optimization Using Abstract Interpretation (and The Ciao System Preprocessor). *Science of Computer Programming*, 58(1–2):115–140, 2005. Available at <https://cliplab.org/papers/ciaopp-sas03-journal-scp.pdf>.
- 2 F. Bueno, P. Lopez-Garcia, J. F. Morales, G. Puebla, and M. V. Hermenegildo. The Ciao Program Preprocessor. Technical Report CLIP-2/2025, Technical University of Madrid (UPM), Facultad de Informática, 28660 Boadilla del Monte, Madrid, Spain, June 2025. Available at <https://ciao-lang.org/ciao/build/doc/ciaopp.html/>.
- 3 I. Garcia-Contreras, D. Ferreira, J. F. Morales, F. Bueno, G. Puebla, P. Lopez-Garcia, and M. V. Hermenegildo. CiaoPP Tutorials. Technical Report CLIP-3/2025, Technical University of Madrid (UPM), Facultad de Informática, 28660 Boadilla del Monte, Madrid, Spain, January 2025. Available at https://ciao-lang.org/ciao/build/doc/ciaopp_tutorials.html.
- 4 The Ciao Developers. The Ciao Playground. Available at <https://ciao-lang.org/playground/>.
- 5 Clark Barrett, Pascal Fontaine, Cesare Tinelli. The Satisfiability Modulo Theories Library (SMT-LIB). www.SMT-LIB.org, 2016.
- 6 C. A. R. Hoare. The Verifying Compiler: A Grand Challenge for Computing Research. In: *Perspectives of Systems Informatics, 5th International Andrei Ershov Memorial Conference, PSI 2003, Akademgorodok, Novosibirsk, Russia, July 9-12, 2003, Revised Papers*. Ed. by Manfred Broy and Alexandre V. Zamulin. Lecture Notes in Computer Science, volume 2890, pages 1–12, Springer, 2003. 10.1007/978-3-540-39866-0_1
- 7 Tony Hoare, Jayadev Misra, Gary T. Leavens, Natarajan Shankar. The Verified Software Initiative: A Manifesto. In: *Theories of Programming: The Life and Works of Tony Hoare*. Ed. by Cliff B. Jones and Jayadev Misra. ACM Books, volume 39, pages 81–92, ACM / Morgan & Claypool, 2021. 10.1145/3477355.3477361

- 8 Helmut Seidl, Ralf Vogler. Three Improvements to the Top-Down Solver. In: *Proceedings of the 20th International Symposium on Principles and Practice of Declarative Programming, PPDP 2018, Frankfurt am Main, Germany, September 03-05, 2018*. Ed. by David Sabel and Peter Thiemann. Pages 21:1–21:14, ACM, 2018. 10.1145/3236950.3236967
- 9 Daniel Baier, Dirk Beyer, Karlheinz Friedberger. JavaSMT 3: Interacting with SMT Solvers in Java. In: *Computer Aided Verification – 33rd International Conference, CAV 2021, Virtual Event, July 20-23, 2021, Proceedings, Part II*. Ed. by Alexandra Silva and K. Rustan M. Leino. Lecture Notes in Computer Science, volume 12760, pages 195–208, Springer, 2021. 10.1007/978-3-030-81688-9_9
- 10 Gidon Ernst, Paula Herber, Marieke Huisman, Mattias Ulbrich. SpecifyThis Bridging Gaps Between Program Specification Paradigms: Track Introduction. In: *Leveraging Applications of Formal Methods, Verification and Validation. Specification and Verification – 12th International Symposium, ISOFA 2024, Crete, Greece, October 27-31, 2024, Proceedings, Part III*. Ed. by Tiziana Margaria and Bernhard Steffen. Lecture Notes in Computer Science, volume 15221, pages 3–7, Springer, 2024. 10.1007/978-3-031-75380-0_1
- 11 Dirk Beyer, Matthias Dangel, Daniel Dietsch, Matthias Heizmann, Thomas Lemberger, Michael Tautschnig. Verification Witnesses. *ACM Trans. Softw. Eng. Methodol.*, 31(4): 57:1–57:69, 2022. 10.1145/3477579
- 12 Bertrand Jeannet, Antoine Miné. Apron: A Library of Numerical Abstract Domains for Static Analysis. In: *Computer Aided Verification, 21st International Conference, CAV 2009, Grenoble, France, June 26 – July 2, 2009. Proceedings*. Ed. by Ahmed Bouajjani and Oded Maler. Lecture Notes in Computer Science, volume 5643, pages 661–667, Springer, 2009. 10.1007/978-3-642-02658-4_52
- 13 Roberto Bagnara, Patricia M. Hill, Enea Zaffanella. The Parma Polyhedra Library: Toward a complete set of numerical abstractions for the analysis and verification of hardware and software systems. *Sci. Comput. Program.*, 72(1-2): 3–21, 2008. 10.1016/j.scico.2007.08.001
- 14 Gagandeep Singh, Markus Püschel, Martin T. Vechev. Fast polyhedra abstract domain. In: *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages, POPL 2017, Paris, France, January 18-20, 2017*. Ed. by Giuseppe Castagna and Andrew D. Gordon. Pages 46–59, ACM, 2017. 10.1145/3009837.3009885
- 15 Julian Erhard, Manuel Bentele, Matthias Heizmann, Dominik Klumpp, Simmo Saan, Frank Schüssele, Michael Schwarz, Helmut Seidl, Sarah Tilscher, Vesal Vojdani. Correctness Witnesses for Concurrent Programs: Bridging the Semantic Divide with Ghosts. In: *Verification, Model Checking, and Abstract Interpretation – 26th International Conference, VMCAI 2025, Denver, CO, USA, January 20-21, 2025, Proceedings, Part I*. Ed. by Shankaranarayanan Krishna, Sriram Sankaranarayanan, and Ashutosh Trivedi. Lecture Notes in Computer Science, volume 15529, pages 74–100, Springer, 2025. 10.1007/978-3-031-82700-6_4
- 16 Anna Becchi, Enea Zaffanella. PPLite: Zero-overhead encoding of NNC polyhedra. *Inf. Comput.*, 275: 104620, 2020. 10.1016/j.ic.2020.104620
- 17 Kristin Yvonne Rozier, Rohit Dureja, Ahmed Irfan, Chris Johannsen, Karthik Nukala, Natarajan Shankar, Cesare Tinelli, Moshe Y. Vardi. MoXI: An Intermediate Language for Symbolic Model Checking. In: *Model Checking Software – 30th International Symposium, SPIN 2024, Luxembourg City, Luxembourg, April 8-9, 2024, Proceedings*. Ed. by Thomas Neele and Anton Wijs. Lecture Notes in Computer Science, volume 14624, pages 26–46, Springer, 2024. 10.1007/978-3-031-66149-5_2
- 18 Marco Elio Gustavo Gario, Andrea Micheli. PySMT: a Solver-Agnostic Library for Fast Prototyping of SMT-Based Algorithms. In: *Proceedings of the 13th International Workshop on Satisfiability Modulo Theories (SMT)*, 2015.

- 19 Kalyan Muthukumar, Manuel V. Hermenegildo. Compile-Time Derivation of Variable Dependency Using Abstract Interpretation. *J. Log. Program.*, 13(2&3): 315–347, 1992. 10.1016/0743-1066(92)90035-2
- 20 Paulína Ayaziová, Dirk Beyer, Marian Lingsch Rosenfeld, Martin Spiessl, Jan Strejček. Software Verification Witnesses 2.0. In: *Model Checking Software – 30th International Symposium, SPIN 2024, Luxembourg City, Luxembourg, April 8-9, 2024, Proceedings*. Ed. by Thomas Neele and Anton Wijs. Lecture Notes in Computer Science, volume 14624, pages 184–203, Springer, 2024. 10.1007/978-3-031-66149-5_11
- 21 Arie Gurfinkel, Jorge A. Navas. Abstract Interpretation of LLVM with a Region-Based Memory Model. In: *Software Verification – 13th International Conference, VSTTE 2021, New Haven, CT, USA, October 18-19, 2021, and 14th International Workshop, NSV 2021, Los Angeles, CA, USA, July 18-19, 2021, Revised Selected Papers*. Ed. by Roderick Bloem, Rayna Dimitrova, Chuchu Fan, and Natasha Sharygina. Lecture Notes in Computer Science, volume 13124, pages 122–144, Springer, 2021. 10.1007/978-3-030-95561-8_8
- 22 Arie Gurfinkel, Temesghen Kahsai, Anvesh Komuravelli, Jorge A. Navas. The SeaHorn Verification Framework. In: *Computer Aided Verification – 27th International Conference, CAV 2015, San Francisco, CA, USA, July 18-24, 2015, Proceedings, Part I*. Ed. by Daniel Kroening and Corina S. Pasareanu. Lecture Notes in Computer Science, volume 9206, pages 343–361, Springer, 2015. 10.1007/978-3-319-21690-4_20
- 23 Elazar Gershuni, Nadav Amit, Arie Gurfinkel, Nina Narodytska, Jorge A. Navas, Noam Rinetzky, Leonid Ryzhyk, Mooly Sagiv. Simple and precise static analysis of untrusted Linux kernel extensions. In: *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2019, Phoenix, AZ, USA, June 22-26, 2019*. Ed. by Kathryn S. McKinley and Kathleen Fisher. Pages 1069–1084, ACM, 2019. 10.1145/3314221.3314590
- 24 Maria Christakis, Hasan Ferit Eniser, Holger Hermanns, Jörg Hoffmann, Yugesh Kothari, Jianlin Li, Jorge A. Navas, Valentin Wüstholtz. Automated Safety Verification of Programs Invoking Neural Networks. In: *Computer Aided Verification – 33rd International Conference, CAV 2021, Virtual Event, July 20-23, 2021, Proceedings, Part I*. Ed. by Alexandra Silva and K. Rustan M. Leino. Lecture Notes in Computer Science, volume 12759, pages 201–224, Springer, 2021. 10.1007/978-3-030-81685-8_9
- 25 Vincenzo Arceri, Greta Dolcetti, Enea Zaffanella. Speeding up Static Analysis with the Split Operator. In: *Proceedings of the 12th ACM SIGPLAN International Workshop on the State Of the Art in Program Analysis, SOAP 2023, Orlando, FL, USA, 17 June 2023*. Ed. by Pietro Ferrara and Liana Hadarean. Pages 14–19, ACM, 2023. 10.1145/3589250.3596141
- 26 Guangsheng Fan, Liqian Chen, Banghu Yin, Wenyu Zhang, Peisen Yao, Ji Wang. Program Analysis Combining Generalized Bit-Level and Word-Level Abstractions. *Proc. ACM Softw. Eng.*, 2(ISSTA): 663–685, 2025. 10.1145/3728905
- 27 Zsófia Ádám, Paulína Ayaziová, Levente Bajczi, Dirk Beyer, Marek Jankola, Marian Lingsch-Rosenfeld, Jan Strejček. Non-termination Witnesses and Their Validation. 2025.
- 28 Patrick Baudin, Pascal Cuoq, Jean-Christophe Filliâtre, Claude Marché, Benjamin Monate, Yannick Moy, Virgile Prevosto. ANSI/ISO C Specification Language Version 1.22. 2025.
- 29 OASIS Open. Static Analysis Results Interchange Format (SARIF) Version 2.1.0. OASIS Standard, April 2020. <https://docs.oasis-open.org/sarif/sarif/v2.1.0/sarif-v2.1.0.html>
- 30 Dirk Beyer, Gidon Ernst, Martin Jonáš, Marian Lingsch-Rosenfeld. SV-LIB: A Standard Exchange Format for Software-Verification Tasks. Technical Report, 2025. https://www.sosy-lab.org/research/pub/2025-TR.SV-LIB_A_Standard_Exchange_Format_for_Software-Verification_Tasks.pdf

- 31 Jorge Navas, Arie Gurfinkel. Clam: Static Analyzer for LLVM Bitcode based on Abstract Interpretation. Version dev14, 2024. <https://github.com/seahorn/clam>
- 32 Pierre Lermusiaux, Benoît Montagu. Detection of Uncaught Exceptions in Functional Programs by Abstract Interpretation. In: *Programming Languages and Systems – 33rd European Symposium on Programming, ESOP 2024, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2024, Luxembourg City, Luxembourg, April 6-11, 2024, Proceedings, Part II*. Ed. by Stephanie Weirich. Lecture Notes in Computer Science, volume 14577, pages 391–420, Springer, 2024. 10.1007/978-3-031-57267-8_15
- 33 Simmo Saan, Julian Erhard, Michael Schwarz, Stanimir Bozhilov, Karoliine Holter, Sarah Tilscher, Vesal Vojdani, Helmut Seidl. Goblint: Abstract Interpretation for Memory Safety and Termination – (Competition Contribution). In: *Tools and Algorithms for the Construction and Analysis of Systems – 30th International Conference, TACAS 2024, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2024, Luxembourg City, Luxembourg, April 6-11, 2024, Proceedings, Part III*. Ed. by Bernd Finkbeiner and Laura Kovács. Lecture Notes in Computer Science, volume 14572, pages 381–386, Springer, 2024. 10.1007/978-3-031-57256-2_25
- 34 Simmo Saan, Michael Schwarz, Julian Erhard, Sarah Tilscher, Karoliine Holter, Ralf Vogler, Kalmer Apinis, Vesal Vojdani. Goblint. <https://github.com/goblint/analyzer>. 10.5281/zenodo.5735006
- 35 Antoine Miné, Abdelraouf Ouadjaout, Matthieu Journault, Raphaël Monat, Francesco Parolini, Marco Milanese, Jérôme Boillot. Mopsa. Version 1.2, 2025. <https://gitlab.com/mopsa/mopsa-analyzer>
- 36 Matthieu Journault, Antoine Miné, Raphaël Monat, Abdelraouf Ouadjaout. Combinations of Reusable Abstract Domains for a Multilingual Static Analyzer. In: *Verified Software. Theories, Tools, and Experiments – 11th International Conference, VSTTE 2019, New York City, NY, USA, July 13-14, 2019, Revised Selected Papers*. Ed. by Supratik Chakraborty and Jorge A. Navas. Lecture Notes in Computer Science, volume 12031, pages 1–18, Springer, 2019. 10.1007/978-3-030-41600-3_1
- 37 David Delmas, Abdelraouf Ouadjaout, Antoine Miné. Static Analysis of Endian Portability by Abstract Interpretation. In: *Static Analysis – 28th International Symposium, SAS 2021, Chicago, IL, USA, October 17-19, 2021, Proceedings*. Ed. by Cezara Dragoi, Suvam Mukherjee, and Kedar S. Namjoshi. Lecture Notes in Computer Science, volume 12913, pages 102–123, Springer, 2021. 10.1007/978-3-030-88806-0_5
- 38 David Delmas, Antoine Miné. Analysis of Software Patches Using Numerical Abstract Interpretation. In: *Static Analysis – 26th International Symposium, SAS 2019, Porto, Portugal, October 8-11, 2019, Proceedings*. Ed. by Bor-Yuh Evan Chang. Lecture Notes in Computer Science, volume 11822, pages 225–246, Springer, 2019. 10.1007/978-3-030-32304-2_12
- 39 Francesco Parolini, Antoine Miné. Sound Abstract Nonexploitability Analysis. In: *Verification, Model Checking, and Abstract Interpretation – 25th International Conference, VMCAI 2024, London, United Kingdom, January 15-16, 2024, Proceedings, Part II*. Ed. by Rayna Dimitrova, Ori Lahav, and Sebastian Wolff. Lecture Notes in Computer Science, volume 14500, pages 314–337, Springer, 2024. 10.1007/978-3-031-50521-8_15
- 40 Marco Milanese, Antoine Miné. Under-Approximating Memory Abstractions. In: *Static Analysis – 31st International Symposium, SAS 2024, Pasadena, CA, USA, October 20-22, 2024, Proceedings*. Ed. by Roberto Giacobazzi and Alessandra Gorla. Lecture Notes in Computer Science, volume 14995, pages 300–326, Springer, 2024. 10.1007/978-3-031-74776-2_12
- 41 Raphaël Monat, Abdelraouf Ouadjaout, Antoine Miné. Easing maintenance of academic static analyzers. *Int. J. Softw. Tools Technol. Transf.*, 26(6): 673–686, 2024. 10.1007/s10009-024-00770-1

- 42 Matthieu Journault, Antoine Miné, Abdelraouf Ouadjaout. Modular Static Analysis of String Manipulations in C Programs. In: *Static Analysis – 25th International Symposium, SAS 2018, Freiburg, Germany, August 29-31, 2018, Proceedings*. Ed. by Andreas Podelski. Lecture Notes in Computer Science, volume 11002, pages 243–262, Springer, 2018. 10.1007/978-3-319-99725-4_16
- 43 Raphaël Monat. Try-Mopsa: Relational Static Analysis in Your Pocket. ArXiv, abs/2509.13128, 2025. arxiv.org/abs/2509.13128
- 44 Abdelraouf Ouadjaout, Antoine Miné. A Library Modeling Language for the Static Analysis of C Programs. In: *Static Analysis – 27th International Symposium, SAS 2020, Virtual Event, November 18-20, 2020, Proceedings*. Ed. by David Pichardie and Mihaela Sighireanu. Lecture Notes in Computer Science, volume 12389, pages 223–247, Springer, 2020. 10.1007/978-3-030-65474-0_11
- 45 Raphaël Monat, Abdelraouf Ouadjaout, Antoine Miné. Static Type Analysis by Abstract Interpretation of Python Programs. In: *34th European Conference on Object-Oriented Programming, ECOOP 2020, November 15-17, 2020, Berlin, Germany (Virtual Conference)*. Ed. by Robert Hirschfeld and Tobias Pape. LIPIcs, volume 166, pages 17:1–17:29, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. 10.4230/LIPIcs.ECOOP.2020.17
- 46 Raphaël Monat, Abdelraouf Ouadjaout, Antoine Miné. A Multilanguage Static Analysis of Python Programs with Native C Extensions. In: *Static Analysis – 28th International Symposium, SAS 2021, Chicago, IL, USA, October 17-19, 2021, Proceedings*. Ed. by Cezara Dragoi, Suvam Mukherjee, and Kedar S. Namjoshi. Lecture Notes in Computer Science, volume 12913, pages 323–345, Springer, 2021. 10.1007/978-3-030-88806-0_16
- 47 Milla Valnet, Raphaël Monat, Antoine Miné. Compositional Static Value Analysis for Higher-Order Numerical Programs (Artifact). *Dagstuhl Artifacts Ser.*, 11(2): 5:1–5:5, 2025. 10.4230/DARTS.11.2.5
- 48 Zhongyi Wang, Linyu Yang, Mingshuai Chen, Yixuan Bu, Zhiyang Li, Qiuye Wang, Shengchao Qin, Xiao Yi, Jianwei Yin. Parf: Adaptive Parameter Refining for Abstract Interpretation. In: *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 – November 1, 2024*. Ed. by Vladimir Filkov, Baishakhi Ray, and Minghui Zhou. Pages 1082–1093, ACM, 2024. 10.1145/3691620.3695487
- 49 Markus Fleischmann, David Kaundlstorfer, Anastasia Isychev, Valentin Wüstholtz, Maria Christakis. Constraint-Based Test Oracles for Program Analyzers. In: *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 – November 1, 2024*. Ed. by Vladimir Filkov, Baishakhi Ray, and Minghui Zhou. Pages 344–355, ACM, 2024. 10.1145/3691620.3695035
- 50 David Kaundlstorfer, Anastasia Isychev, Valentin Wüstholtz, Maria Christakis. Interrogation Testing of Program Analyzers for Soundness and Precision Issues. In: *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 – November 1, 2024*. Ed. by Vladimir Filkov, Baishakhi Ray, and Minghui Zhou. Pages 319–330, ACM, 2024. 10.1145/3691620.3695034

4.3 Explainability in Static Analysis

Discussion Leader: Roberto Bagnara, University of Parma

Attendees

Roberto Bagnara, Valentin Barbazo, Anna Becchi, Patrick Cousot, Kihong Heo, Tim King, Naim Moussaoui-Remil, Jorge Navas.

Agenda

1. Define and shape the topic, and identify possible contributions.
2. Decide on potential next steps.

4.3.1 Preliminary Notes

There is some (limited) literature on the subject. See, e.g., [3] and [1].

4.3.2 Tentative Definition

For a static analysis tool, *explainability* denotes its ability to explain its findings to a given intended audience.

The most typical audience consists of development and QA teams, possibly with different skill sets and needs. This diversity should be considered, even though it was not directly discussed during the meeting.

4.3.3 What Has to Be Explained

Ideally, the tool should, for each of its findings or messages, explain – or at least make it easy for the audience to understand – the following aspects (the list may not be exhaustive):

1. **The bad thing to be avoided and why it is problematic.** Examples: memory leak (undesirable because it can lead to memory exhaustion); misaligned access to memory (undesirable because it is undefined behavior in C and C++).
2. **The specific finding.** Examples: A pointer to allocated memory went out of scope without freeing the associated memory; a pointer cast that may result in a misaligned pointer.
3. **The connection between the finding and the undesired behavior.** Example: If the pointer that went out of scope was the only reference to the allocated memory block, and that block was not deallocated, a memory leak occurs. Similarly, if a cast produces a misaligned pointer that is later dereferenced, undefined behavior occurs.
4. **Whether the finding has a definite or possible nature.** Examples: Is it certain that the deallocation function was not called before the pointer went out of scope? Is it certain that the cast was executed and could indeed result in a misaligned pointer?
5. **How the static analyzer decided to report the finding.** Example: Variable `ptr` received the result of `malloc` at program point p_1 and the returned value is definitely/possibly non-null. There exists a computation path leading to p_2 , where `ptr` goes out of scope. The path is definitely/possibly feasible; `ptr` is definitely/possibly not copied to a memory location that survives `ptr`; the memory pointed to by `ptr` is definitely/possibly not deallocated. Hence, the memory block allocated at p_1 is definitely/possibly leaked.

Some of the above aspects may be trivial or redundant for certain audiences; in those cases, a reference to relevant documentation may suffice. Furthermore:

- most tools already provide the basic information (points 1–3);
- high-quality tools typically distinguish well between definite and possible findings (point 4);

Point 5 – explaining why a finding is reported – is particularly challenging and where further research is most needed.

4.3.4 Discussion Highlights

- **Iterated forward + backward analysis** (suggested by Cousot and others): useful for excluding false positives, but does not inherently explain the remaining positives.
- **Model checking / symbolic execution / automatic test-case generation**: can filter out false positives and also produce a witness, arguably the best form of explanation for any audience, at least for those cases where the witness is of manageable size.
- **Tracking dependencies during analysis** (suggested by Bagnara): record dependencies of abstract values alongside the values themselves, allowing users to understand the source of imprecision and refine the analysis through annotations or re-execution.
Example: Here we have the addition of $a + b$ for which we cannot exclude an arithmetic overflow. Where do the (too large) intervals for a and b come from? They come from function parameters x_1 and x_2 and widening at program point p_1 . If we present this information to the user, they may add assertions about the possible values of x_1 and x_2 . We rerun the analysis (iterated forward + backward) and maybe the problem goes away. If it does not go away, maybe the issue is real or maybe the problem is in the widening at program point p_1 . What can the user do with the information that “maybe the problem is the widening at program point p_1 ”? Adding an annotation at program point p_1 saying “tool, try harder here”?
- **Widening-based idea** (Cousot): record the delta used during widening. When an error is reached again, retry with a different widening delta.
- **Statistical prioritization** (Heo): use statistical analysis of invariants to estimate whether a warning is more or less likely to be a true positive, effectively measuring analysis precision.
- **Meta-analysis / “analysis of the analysis”** (Cousot): track dependencies and probabilistic data to record the analyzer’s internal reasoning, recalling the ideas of A²I (i.e., analysis of the analysis) [2].

4.3.5 Possible Action Items

1. Review the available literature.
2. Reflect further on the discussed approaches.
3. Consider the possible next steps (e.g., establishing a working group for the purpose of writing a paper, or organizing a new Dagstuhl Seminar on the topic).

References

- 1 Marcus Nachtigall, Lisa Nguyen Quang Do, Eric Bodden. Explaining Static Analysis – A Perspective. In: *34th IEEE/ACM International Conference on Automated Software Engineering – Workshop (ASEW)*, pages 29–32, 2019. 10.1109/ASEW.2019.00023
- 2 Patrick Cousot, Roberto Giacobazzi, Francesco Ranzato. A²I: Abstract² Interpretation. *Proc. ACM Program. Lang.*, 3(POPL): 42:1–42:31, January 2019. 10.1145/3290355
- 3 Eric Bodden, Lisa Nguyen Quang Do. Explainable Static Analysis. In: *Software Engineering und Software Management 2018*, pages 205–208. Gesellschaft für Informatik, Bonn, 2018. ISBN 978-3-88579-673-2.

4.4 Combination of LLMs and Static Analysis

Discussion Leader: Greta Dolcetti, Ca' Foscari University of Venice

The adoption of Large Language Models (LLMs) and static analysis together is reshaping how developers interact with code analysis tools [1]. This emerging field offers compelling opportunities to enhance verification workflows, democratize complex analysis techniques, and bridge the gap between technical analysis outputs and developer understanding.

Our discussion group explored how to enhance the developer and user experience through natural language and how hybrid approaches can combine the strengths of these two areas.

4.4.1 Enhancing Developer and User Experience Through Natural Language

Traditional static analysis tools often present barriers to adoption through complex configuration requirements and technical jargon. LLMs might transform this landscape by serving as intuitive natural language interfaces. LLMs might help contextualize findings within specific code contexts. They can transform cryptic error messages into clear explanations [2] that relate directly to the developer's code, making static analysis more accessible to teams with varying levels of expertise. However, LLMs demonstrate notable limitations in certain critical analysis areas where traditional static analysis methods remain superior, highlighting the complementary rather than competitive nature of these technologies. Another promising application lies in automating the creation of custom analysis rules. LLMs can interpret human-readable specifications and generate corresponding analysis rules[3], dramatically reducing the manual effort and difficulty traditionally required for custom static analysis configurations, especially domain-specific requirements or legacy codebases requiring tailored analysis approaches. The integration becomes even more powerful when combined with synthesis techniques. Using frameworks like Syntax-Guided Synthesis (SyGus) [4], LLMs can generate analysis rules for specific patterns that organizations want to detect, then incorporate these rules into feedback systems that allow for human intervention and verification. This creates a collaborative workflow where LLMs handle the initial heavy lifting while human expertise ensures correctness and relevance. While manual verification remains necessary to ensure query correctness, the automated generation significantly accelerates the development cycle.

4.4.2 Hybrid Approaches: Combining Strengths

The most effective applications often emerge from hybrid approaches that leverage both LLM capabilities and traditional static analysis strengths. In computationally expensive analyses such as floating-point analysis, LLMs can provide initial hints to prune the search space, with traditional tools validating the results. This combination addresses scalability challenges while maintaining the correctness guarantees that static analysis provides. Multi-model strategies offer another promising direction. Running multiple LLMs on the same repository and checking for agreement increases confidence in results, as different models may excel in different contexts. This ensemble approach can improve overall accuracy while providing insights into the reliability of specific findings. Fine-tuning presents additional opportunities for improvement. Domain-specific training can enhance LLM performance in static analysis tasks, though this approach requires significant domain expertise and carefully curated training data. LLMs also show promise in addressing one of static analysis's persistent challenges: false positive management [5]. They can assist in triaging potential false positives, grouping findings with similar root causes, and generating counterexamples that help analysts understand analysis results. This capability can significantly reduce the manual effort required to derive actionable insights from static analysis outputs.

References

- 1 VN Ignatyev, NV Shimchik, DD Panov, AA Mitrofanov. Large language models in source code static analysis. 2024 Ivannikov Memorial Workshop (IVMEM), pages 28–35, IEEE, 2024.
- 2 John S. Y. Lee, Fengkai Liu, Tianyuan Cai. Code Debugging with LLM-Generated Explanations of Programming Error Messages. 2024 IEEE 13th International Conference on Engineering Education (ICEED), pages 1–5, IEEE, 2024. 10.1109/ICEED62316.2024.10923833
- 3 Chenyuan Yang, Zijie Zhao, Zichen Xie, Haoyu Li, Lingming Zhang. KNighter: Transforming Static Analysis with LLM-Synthesized Checkers. Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles, 2025.
- 4 Rajeev Alur, Rastislav Bodik, Garvit Juniwal, Milo M. K. Martin, Mukund Raghothaman, Sanjit A. Seshia, Rishabh Singh, Armando Solar-Lezama, Emina Torlak, Abhishek Udupa. Syntax-guided synthesis. 2013 Formal Methods in Computer-Aided Design, pages 1–8, IEEE, 2013. 10.1109/FMCD.2013.6679385
- 5 Xueying Du, Kai Yu, Chong Wang, Yi Zou, Wentai Deng, Zuoyu Ou, Xin Peng, Lingming Zhang, Yiling Lou. Minimizing False Positives in Static Bug Detection via LLM-Enhanced Path Feasibility Analysis. ArXiv, abs/2506.10322, 2025.

4.5 Shifting Abstract Interpretation Left: Using Abstraction Interpretation Earlier in the Development Process

Discussion Leader: Michael Schwarz, National University of Singapore

Currently, abstract interpretation is often used only after a program has been developed. However, performing static analysis during development could provide significant benefits, ranging from easing development to detecting bugs much earlier or even preventing them altogether.

Our discussion group explored how to “shift left” abstract interpretation, the prerequisites for doing so, and promising future directions. One of the identified key prerequisites is to make abstract interpreters *incremental*, i.e., able to efficiently re-analyze a program after small changes. While there is existing work [3, 1, 2, 4], new execution models prevalent in, e.g., data science notebooks pose novel incremental analysis problems, as the effects of previous versions may linger in the environment. This poses challenges at all levels, ranging from defining a version-aware concrete semantics to designing suitable abstract domains. We also discussed why it may be an advantage to maintain results for several old program versions: It may help recover precision if a catastrophic loss of precision occurs for an intermediate version, or allow an incremental analysis to start from the version most semantically similar to the one to be analyzed. We also observed a relationship between incremental analysis and relational analyses relating multiple program versions. It may, for instance, be helpful to alert a user if, after a change intended to be purely cosmetic, the equivalence of the old and new versions can no longer be established.

Beyond incremental analysis, we discussed how to involve developers more directly in the analysis process. For example, they could supply candidate invariants during development. While one may not want to take such invariants at face value, they can guide the analysis in a sound way, e.g., by acting as widening thresholds [5]. Furthermore, it may be useful to provide immediate feedback to developers, e.g., when a call to a function is the first one for which the analysis cannot prove that the argument is non-null, or when a function argument has a value far outside the range of previous calls. While such situations may not be bugs

in themselves, they can serve as useful hints to the programmer that something may be amiss and may help avoid the introduction of bugs in the first place. Similarly, the analysis may flag instances where precision degrades significantly, prompting the user to write code more amenable to static analysis and thus likely also easier for humans to understand. Such interactions may help shape code in a way similar to style guides, except that they are tied to a more semantic property, namely whether the analyzer can reason about the code precisely.

We also discussed designing a specification language for developers to express their own properties of interest. To reach widespread adoption, such a specification language may follow the example of the popular semantic patching tool **Coccinelle** [6] by leveraging familiar syntax to lower the barrier to entry, perhaps by building on existing mechanisms such as **CodeQL** [7].

One challenge facing work on left-shifting abstract interpretation is the question of how to evaluate such techniques, especially when they involve interactions with users. We concluded that such evaluations may have to involve user studies, which can perhaps best be conducted by collaborating with other communities, such as the HCI community, as few researchers have experience both in sound static analysis and in conducting user studies.

References

- 1 Julian Erhard, Simmo Saan, Sarah Tilscher, Michael Schwarz, Karoliine Holter, Vesal Vojdani, Helmut Seidl. Interactive abstract interpretation: reanalyzing multithreaded C programs for cheap. *International Journal on Software Tools for Technology Transfer*, 26(6): 647–667, 2024. 10.1007/s10009-024-00768-9
- 2 Benno Stein, Bor-Yuh Evan Chang, Manu Sridharan. Demanded abstract interpretation. In: *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pages 282–295, ACM, 2021. 10.1145/3453483.3454044
- 3 Isabel Garcia-Contreras, José F. Morales, Manuel V. Hermenegildo. Incremental and Modular Context-sensitive Analysis. *Theory and Practice of Logic Programming*, 21(2): 211–243, 2021. 10.1017/S1471068420000496
- 4 Mamy Razafintsialonina, David Bühler, Antoine Miné, Valentin Perrelle, Julien Signoles. Reusing Caches and Invariants for Efficient and Sound Incremental Static Analysis (Extended Version). In: *39th European Conference on Object-Oriented Programming (ECOOP 2025)*, Bergen, Norway, 2025.
- 5 Nicolas Halbwachs, Yann-Erick Proy, Patrick Roumanoff. Verification of Real-Time Systems using Linear Relation Analysis. *Formal Methods in System Design*, 11(2): 157–185, 1997. 10.1023/A:1008678014487
- 6 Yoann Padioleau, Julia Lawall, René Rydhof Hansen, Gilles Muller. Documenting and automating collateral evolutions in linux device drivers. In: *Proceedings of the 3rd ACM SIGOPS/EuroSys European Conference on Computer Systems 2008*, pages 247–260, ACM, 2008. 10.1145/1352592.1352618
- 7 Oege De Moor, Mathieu Verbaere, Elnar Hajiyeve, Pavel Avgustinov, Torbjorn Ekman, Neil Ongkingco, Damien Sereni, Julian Tibble. Keynote address: QL for source code analysis. In: *Seventh IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM 2007)*, pages 3–16, IEEE, 2007.

5 Social Activities

Pietro Ferrara (University of Venice, IT), Liana Hadarean (Amazon Web Services – Seattle, US), Jorge Navas (Certora – Seattle, US), and Caterina Urban (INRIA & ENS Paris, FR)
License © Creative Commons BY 4.0 International license
© Pietro Ferrara, Liana Hadarean, Jorge Navas, and Caterina Urban

During the first two days of the seminar, we organized a table football (Figure 1) and a tennis table (Figure 2) tournament. We formed teams of two participants for both tournaments; eight teams participated in each tournament (Table 1 reports the teams). We organized single-elimination tournaments starting with the quarter-finals. Since everybody wanted to play several games, we also organized semi-finals and finals for teams that had lost earlier matches, resulting in a total ordering in the final standings for all the teams. Teams were “randomly” paired using ChatGPT. The results of the two tournaments are reported by Tables 1 and 2. The competition was tough, and, as reported by Figure 1, one table football player got injured after the tournament. We hope that the Dagstuhl staff was able to fix it before the next seminar started.

In the late evenings, we played several board games (Figure 3), but unfortunately, we neither organized tournaments nor kept track of the results.



Figure 1 Table football tournament.



Figure 2 Tennis table tournament.



■ **Figure 3** Board games.

■ **Table 1** Teams.

Fussbal		Tennis table	
Name	Players	Name	Players
LiSA	Arceri, Ferrara	Seahorn	Navas, Wang/Gurfinkel
Function	Urban, Moussaoui	LiSA	Ferrara, Dolcetti
MOPSA	Monat, Mine	MOPSA	Mine', Monat
MemCAD	Barbazo, De Haro	YoYak	Minseok, Kihong
SeaHorn	Delmas, Becchi	Function	Urban, Moussaoui
GIDA	Denaro, Taibi	MemCAD	Barbazo, De Haro
PPL	Bagnara, Dolcetti	Llm	Dinella, Tan
Last minute tea (LMT)	Garcia, Dinella	Aws	Hadarean, Balliu

■ **Table 2** Table football tournament.

Quarter-finals			Semi-finals		
	MemCAD-PPL	12-10	1st-4th	MemCAD-Function	7-10
	Function-MOPSA	10-8	1st-4th	GIDA-LiSA	10-8
	GIDA-LMT	10-5	5th-8th	PPL-MOPSA	10-4
	SeaHorn-LiSA	6-10	5th-8th	LMT-SeaHorn	12-14
Finals			Standing		
1st-2nd	GIDA-Function	10-6	1)	GIDA	
3rd-4th	MemCAD-LiSA	3-10	2)	Function	
5th-6th	PPL-SeaHorn	10-5	3)	LiSA	
7th-8th	MOPSA-LMT	10-8	4)	MemCAD	
			5)	PPL	
			6)	SeaHorn	
			7)	MOPSA	
			8)	LMT	

■ **Table 3** Tennis table tournament.

Quarter-finals			Semi-finals		
	Seahorn-Llm	13-21	1st-4th	Llm-Aws	11-21
	Aws-YoYak	21-14	1st-4th	MOPSA-MemCAD	15-21
	LiSA-MOPSA	14-21	5th-8th	Seahorn-YoYak	21-6
	Function-MemCAD	12-21	5th-8th	LiSA-Function	21-19
Finals			Standing		
1st-2nd	Aws-Memcad	20-22	1)	MemCAD	
3rd-4th	Llm-MOPSA	21-19	2)	Aws	
5th-6th	Seahorn-LiSA	21-12	3)	Llm	
7th-8th	YoYak-Function	21-22	4)	MOPSA	
			5)	Seahorn	
			6)	LiSA	
			7)	Function	
			8)	Yoyak	

Participants

- Vincenzo Arceri
University of Parma, IT
- Roberto Bagnara
University of Parma, IT
- Musard Balliu
KTH Royal Institute of
Technology – Stockholm, SE
- Valentin Barbazo
ENS, PSL University – Paris, FR
- Anna Becchi
USI – Lugano, CH
- Dirk Beyer
LMU München, DE
- Patrick Cousot
New York University, US
- Eva Darulova
Uppsala University, SE
- Charles De Haro
ENS, PSL University – Paris, FR
- David Delmas
Airbus – Toulouse, FR
- Giovanni Denaro
University of Milano-Bicocca, IT
- Werner Dietl
University of Waterloo, CA
- Elizabeth Dinella
Bryn Mawr College, US
- Greta Dolcetti
University of Venice, IT
- Pietro Ferrara
University of Venice, IT
- Isabel Garcia-Contreras
Black Duck – Calgary, CA
- Elisa Gonzalez Boix
VU – Brussels, BE
- Arie Gurfinkel
University of Waterloo, CA
- Liana Hadarean
Amazon Web Services –
Seattle, US
- Kihong Heo
KAIST – Daejeon, KR
- Manuel Hermenegildo
IMDEA Software Institute –
Madrid, ES
- Minseok Jeon
DGIST Institute of Science &
Technology – Daegu, KR
- Tim King
Amazon Web Services –
Santa Clara, US
- Valentina Lenarduzzi
University of Oulu, FI
- Debasmita Lohar
KIT – Karlsruher Institut für
Technologie, DE
- Antoine Miné
Sorbonne University – Paris, FR
- Raphaël Monat
INRIA Lille, FR
- Naïm Moussaoui Remil
ENS, PSL University – Paris, FR
- Jorge Navas
Certora – Seattle, US
- Guido Salvaneschi
Universität St. Gallen, CH
- Michael Schwarz
National University of
Singapore, SG
- Helmut Seidl
TU München – Garching, DE
- Davide Taibi
University of Southern Denmark –
Odense, DK
- Tian Tan
Nanjing University, CN
participant Caterina Urban
INRIA & ENS Paris, FR
- Jingbo Wang
Purdue University –
West Lafayette, US



Cognitive Sensing and Interaction

Thomas Kosch^{*1}, Kai Kunze^{*2}, Christina Schneegass^{*3}, and Thad Starner^{*4}

1 HU Berlin, DE. thomaskosch90@gmail.com

2 Keio University – Yokohama, JP. kai@kmd.keio.ac.jp

3 TU Delft, NL. c.schneegass@tudelft.nl

4 Georgia Institute of Technology – Atlanta, US. thadstarner@gmail.com

Abstract

This report documents the Dagstuhl Seminar “Cognitive Sensing and Interaction” (25422). The report consists of an executive summary, as well as abstracts on talks, working groups, panel discussions, and open problems sessions.

Seminar October 12–15, 2025 – <https://www.dagstuhl.de/25422>

2012 ACM Subject Classification Human-centered computing

Keywords and phrases cognition-awareness, cognitive augmentation, cognitive interaction, physiological interaction, wearable technology

Digital Object Identifier 10.4230/DagRep.15.10.75

1 Executive Summary

Thomas Kosch (HU Berlin, DE)

Kai Kunze (Keio University – Yokohama, JP)

Christina Schneegass (TU Delft, NL)

Thad Starner (Georgia Institute of Technology – Atlanta, US)

License © Creative Commons BY 4.0 International license

© Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner

In Human-Computer Interaction (HCI), cognitive sensing offers valuable insights into the relationship between users and technologies, enabling new forms of interaction. Monitoring human cognitive processes, such as attention, memory, and decision-making, enables technologies to adapt to user needs, provide personalized support, and develop new methods for measuring user experience. Such data streams allow for seamless communication between humans and computing systems. With recent advances in Artificial Intelligence (AI), we observe significant improvement in one of the major hurdles in cognitive interaction – the precise and reliable classification of cognitive states. However, critical challenges remain before cognitive interaction can be effectively used in real-world environments. This Dagstuhl Seminar “Cognitive Sensing and Interaction” (25422) focused on three of those challenges, namely (a) robust cognitive sensing in everyday life, (b) how to augment human cognitive abilities efficiently, (c) how to design ethical, inclusive, reproducible systems for a diverse population. The goals of this seminar was to identify successful practices, develop a research agenda, and initiate the development of policy documents for designing user-centric technologies that perceive, communicate with, and engage with user cognition. We brought together experts in HCI, neuroscience, AI, and psychology to synthesize solutions for overcoming challenges and discuss ways to integrate cognitive sensing and interaction into our everyday lives.

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Cognitive Sensing and Interaction, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 75–104

Editors: Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

This Dagstuhl Seminar, “Cognitive Sensing and Interaction” (25422), examined how cognitive sensing can transition from lab prototypes to everyday systems that use cognitive interaction. The overarching goal was to discuss the technical feasibility, identify interaction concepts worth pursuing, and determine the necessary governance as sensing technologies mature. The seminar consisted of two keynotes and three days of structured work. The keynotes discussed public-facing use scenarios for personal brain scanners (schools, neurodiversity support, aging) and the move from basic monitoring toward low-power, on-device intelligence enabled by neuromorphic computing. Day 1 mapped sensing modalities and devices to feasibility, including usefulness trade-offs and consolidated metrics, as well as best practices for in-the-wild inference, with an emphasis on multimodal, context-aware interpretation and individualized baselines. Day 2 translated visions into concrete interaction scenarios through rapid prototyping and walk-and-talk ideation, exposing constraints around signal quality, calibration, latency, and how to communicate uncertainty. Day 3 synthesized technical, societal, and legal risks into actionable policy directions and stakeholder responsibilities. The seminar outcomes include an initial framework connecting signals to cognitive constructs and interaction mechanisms, a portfolio of grounded scenarios and design tensions, and a governance agenda centered on transparency, user control, minimal retention, and accountable evaluation beyond accuracy.

2 Table of Contents

Executive Summary

<i>Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner</i>	75
--	----

Overview of Talks

Smart Sensing and Visualization of Human Mobility Experience <i>Esther Bosch</i>	79
Beyond Measurement: Designing Cognitive and Digital Technologies that Support Work and Wellbeing <i>Anna Cox</i>	80
Reinventing Human Potential through Cognitive Augmentations <i>Passant Elagroudy</i>	80
The Future of Wearable Neurotechnology: From Sensing to Neuromorphic Intelligence <i>Mustafa Hamada</i>	81
JIZAI Coding: Embodied, Affective, and Autonomous Programming for Inclusive Creators <i>Masahiko Inami</i>	82
Long-Term Cognitive Augmentation <i>Shoya Ishimaru</i>	82
Open-source neural wearables for the future of work <i>Michael Knierim</i>	83
Estimating Human's Cognitive Situation using AI and CV <i>Hideki Koike</i>	84
Artificial Intelligence for Amplifying Human Perception <i>Thomas Kosch</i>	85
Reflexive Wearables: Turning Physiological Sensing into Human Insight <i>Kai Kunze</i>	86
Inferring Cognition through Oculo-metrics: From Laboratory Research to Real-World Applications <i>Diako Mardanbegi</i>	86
Assistive Augmentation: Fundamentally Transforming Human Ability <i>Suranga Nanayakkara</i>	87
Understanding Cognitive Augmentation through the Lenses of Experience, Agency, and Ethics <i>Jasmin Niess</i>	88
AI-Powered Human Augmentation in the Workplace <i>Evangelos Niforatos</i>	89
From Perception to Shared Cognition: Leveraging Egocentric Physiological and Behavioral Cues for Empathic Remote Collaboration <i>Prasanth Sasikumar</i>	90

Designing Cognition-Aware Interfaces for Neurotechnology	
<i>Christina Schneegass</i>	91
Neuroadaptive and Embodied Interfaces for Cognitive Ergonomics	
<i>Benjamin Tag</i>	92
Beyond Enhancement: Cognitive Sensing Futures	
<i>Steeven Villa</i>	93
AI-Augmented Cognition and Co-Experience in Extended Reality	
<i>Robin Welsch</i>	94
The Future of Cognitive Personal Informatics	
<i>Max L. Wilson</i>	95
Co-designing and Personal Fabrication of Assistive Technologies	
<i>Anusha Withana</i>	96
Working groups	
Envisioning Novel Cognitive Interaction Scenarios	
<i>Thomas Kosch</i>	97
Metrics and Best Practices for Cognitive Sensing	
<i>Thomas Kosch</i>	97
Prototyping Cognitive Interaction Scenarios	
<i>Kai Kunze and Thad Starner</i>	98
Policies for Cognitive Sensing and Cognitive Interaction	
<i>Jasmin Niess and Thomas Kosch</i>	98
Feasibility of Cognitive Sensing Methods in Everyday Life	
<i>Christina Schneegass</i>	100
Vision for Cognitive Sensing in 5, 15, and 50 Years	
<i>Steeven Villa</i>	100
Developing a Framework for Cognitive Interaction	
<i>Max L. Wilson</i>	101
Panel discussions	
The Future of Wearable Neurotech: From Sensing to Neuromorphic Intelligence	
<i>Mustafa Hamada</i>	101
Our Future Living Scenarios with Personal Brain Scanners	
<i>Max L. Wilson</i>	102
Open problems	
Day 1 Synopsis	
<i>Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner</i>	102
Day 2 Synopsis	
<i>Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner</i>	103
Day 3 Synopsis	
<i>Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner</i>	103
Acknowledgments	103
Participants	104

3 Overview of Talks

3.1 Smart Sensing and Visualization of Human Mobility Experience

Esther Bosch (DLR – Braunschweig, DE)

License  Creative Commons BY 4.0 International license
© Esther Bosch

My research investigates how human cognitive, emotional, and experiential states can be measured, understood, and visualized across diverse mobility contexts. Whether traveling by public transport [1], driving a car [2], operating a train [3], or cycling through the city, people continuously experience the mobility system not only as infrastructure but as lived experience. Capturing these experiences systematically offers a powerful way to identify where change in the mobility system is most urgent. To this end, I combine valid and reliable subjective measures of travel experience with physiological and contextual data. These include signals from the cardiovascular system, facial expressions, and electrodermal responses, alongside real-world context such as location, time, and traffic conditions. By linking subjective, contextual and physiological indicators, my work aims to model and visualize travel experience dynamically, enabling transport providers, cities, and communities to see mobility from the traveler's perspective [4]. The broader goal is to make human experience a measurable and actionable component of mobility planning, guiding decisions toward more sustainable and user-centered systems. In the long term, I envision experience measurement that requires minimal user input, grounded in a deep understanding of the relationship between subjective experience and physiological signals. This approach bridges human factors research, psychophysiology, and mobility data science to make the invisible dimensions of travel, such as goal-relevant hindrances, experience compared to expectations, and overall traveler state, visible and usable for design and policy.

References

- 1 E. Bosch, A. R. Luther, and K. Ihme. *Travel experience in public transport: Experience sampling and cardiac activity data for spatial analysis*. *Scientific Data* 12(1), 633, 2025.
- 2 C. Brandebusemeyer, K. Ihme, and E. Bosch. *Travelers' information need in automated vehicles – a psychophysiological analysis*. In *2022 Human-Centered Cognitive Systems (HCCS)*, pp. 1–6. IEEE, December 2022.
- 3 E. J. Bosch, D. Schackmann, S. Hoyer, W. Kilian, S. Schwanitz, and A. Hamann. *Multi-Sensor Vigilance Detection: Sensor Reliability and Comfort in Automated Train Operations*. In *6th German Conference on Rail Human Factors*, 2025.
- 4 E. Bosch, M. Scholz, and K. Ihme. *Experience Atlas: An interactive real-time travel experience visualization*. In *Proceedings of the Transportation Research Symposium 2025*. Rotterdam, Netherlands, 2025.

3.2 Beyond Measurement: Designing Cognitive and Digital Technologies that Support Work and Wellbeing

Anna Cox (University College London, GB)

License  Creative Commons BY 4.0 International license
© Anna Cox

My research explores how digital technologies can be designed and used to support both work and wellbeing. I am particularly interested in the bidirectional relationship between productivity and health: how the very technologies that enable us to work, such as communication, planning, and collaboration tools, can also be harnessed to promote wellbeing, and how the technologies we use to manage our health and wellbeing can, in turn, enhance our effectiveness at work. A major interest is the use of psychological theory to inform the design and evaluation of these digital systems. By grounding technology design in established theories of motivation, self-regulation, and recovery, I aim to deepen our understanding of how people interact with these systems and how technology can be shaped to foster thriving rather than overwork. In related work, I have examined how digital systems for measuring and evaluating work can shape experiences of productivity and control, particularly in remote and hybrid contexts. For example, in *Remote Work, Work Measurement and the State of Work Research in Human-Centred Computing* (Gould et al., 2023), we argued that the expansion of “bossware” and digital surveillance during the pandemic highlights the conceptual limits of technology-mediated understandings of work, and the need for more inclusive, representative approaches to future-of-work research. Within the context of this Dagstuhl Seminar, my interests lie in how insights from psychology and human-computer interaction can inform the responsible development of cognitive and physiological sensing technologies for everyday work. This includes questions of boundary management, trust, and ethical design, ensuring that systems which sense or augment human cognition do so in ways that respect autonomy, diversity, and wellbeing. My contribution to the seminar focused on connecting insights from psychology and HCI to the design of ethical, inclusive cognitive interaction systems.

References

- 1 S. J. J. Gould, A. Rudnicka, D. Cook, M. E. Cecchinato, J. W. Newbold, and A. L. Cox. *Remote Work, Work Measurement and the State of Work Research in Human-Centred Computing*. *Interacting with Computers*, 2023.

3.3 Reinventing Human Potential through Cognitive Augmentations

Passant Elagroudy (DFKI – Kaiserslautern, DE)

License  Creative Commons BY 4.0 International license
© Passant Elagroudy

My research investigates how we can augment human memory and research cognition. A major thread in my work focuses on cognitive alterations – from memory augmentation to attenuation and implantation – enabled by ubiquitous and immersive technologies [1, 2]. I previously explored how technology can act as a partner in remembering – helping people to reconstruct, reinterpret, or even reform their past experiences. Through studies of visual lifelogs, memory cues, and privacy-preserving prostheses, I examine how digital traces interact with our autobiographical memory and sense of self in real life or in virtual

environments [3], revealing both the empowering and altering potentials of “smart” memory systems. More recently, I have turned to Generative AI as a cognitive amplifier for researchers. In collaboration with international HCI researchers, I examine how Large Whatever Models (LWMs) can transform research cycles, automate synthesis, and preserve ubiquitous research artifacts [4, 5]. I envision generative systems as active participants in scientific reasoning and collective memory, capable of extending our cognitive reach and accelerating discovery. Working at the intersection of human-computer interaction, psychology, and AI, I aim to establish hybrid systems that not only augment human memory but also reshape how we conduct, recall, and evolve knowledge. My goal is to design technologies that amplify reflection without erasing human agency – tools that help us remember, reason, and reimagine together.

References

- 1 Passant Elagroudy, Rufat Rzayev, Tonja-Katrin Machulla, Huy Viet Le, Tilman Dingler, Lars Lischke, Sarah Clinch, Geoffrey Ward, and Albrecht Schmidt. *Pixel Memories: Do Lifelog Summaries Fail to Enhance Memory but Offer Privacy-Aware Memory Assessments?*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2025.
- 2 Passant Elagroudy, Mohamed Khamis, Florian Mathis, Diana Irmscher, Ekta Sood, Andreas Bulling, and Albrecht Schmidt. *Impact of Privacy Protection Methods of Lifelogs on Remembered Memories*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2023.
- 3 Amal Yassien, Passant ElAgroupdy, Elhassan Makled, and Slim Abdennadher. *A design space for social presence in VR*. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pp. 1–12, 2020.
- 4 Passant Elagroudy, Jie Li, Kaisa Väänänen, Paul Lukowicz, Hiroshi Ishii, Wendy E. Mackay, Elizabeth F. Churchill, et al. *Transforming HCI research cycles using generative AI and “Large Whatever Models” (LWMs)*. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–5, 2024.
- 5 Albrecht Schmidt, Passant Elagroudy, Fiona Draxler, Frauke Kreuter, and Robin Welsch. *Simulating the human in HCD with ChatGPT: Redesigning interaction design with AI*. *Interactions* 31(1), 24–31, 2024.

3.4 The Future of Wearable Neurotechnology: From Sensing to Neuromorphic Intelligence

Mustafa Hamada (Mendi Innovations – Stockholm, SE)

License © Creative Commons BY 4.0 International license
© Mustafa Hamada

My work explores how neuroscience, behavioral design, and artificial intelligence can be combined to create technologies that advance cognitive health and mental wellbeing. As Chief Product and Science Officer at Mendi Innovations (Sweden), I focus on translating neuroimaging insights into accessible neurofeedback tools, particularly through fNIRS-based brain training for attention and self-regulation. As Honorary Fellow of Neurotechnology at the University of Nottingham’s Brain Data Group, I contribute to research that spans HCI, game design, dynamic system theory, and adaptive neurotechnology, using fNIRS data analysis to study engagement and cognitive load. In parallel, I serve as an Advisor to the REACT project (Responsible AI), that aims at developing responsible, adaptive robotic

systems for complex healthcare environments and co-lead an IEEE workstream developing ethical frameworks for mental health technologies. I am also involved in work investigating how at-edge neuromorphic computing and EEG-based sensing can enable low-power, real-time interpretation of neural signals in wearable systems to predict epilepsy seizure onset. Across these domains, my goal is to advance ethical, personalized, and human-centered neurotechnology that meaningfully integrates with everyday life.

3.5 JIZAI Coding: Embodied, Affective, and Autonomous Programming for Inclusive Creators

Masahiko Inami (University of Tokyo, JP)

License  Creative Commons BY 4.0 International license
© Masahiko Inami

JIZAI Coding is a research initiative that aims to transcend conventional programming paradigms by integrating embodied interaction, affective and physiological sensing, and co-creative AI to empower users – including those with disabilities – to solve their own challenges through code. In contrast to “vibe coding” as a metaphor for loose prompt-driven generation, JIZAI Coding envisions a resonant, adaptive coding partnership between human cognition, bodily states, and generative systems. Drawing upon the framework of Cognitive Sensing and Interaction (Dagstuhl Seminar 25422) and the notion of gut feelings (the body’s tacit cues), this project posits a future where code is not only “written” but felt, synchronized, and lived.

References

- 1 R. W. Picard. *Affective Computing*. MIT Press, 1997.
- 2 S. Salazar. *Searching for gesture and embodiment in live coding*. In *Proceedings of the International Conference on Live Coding*, 2017.
- 3 Adyasha Dash and Kathleen Agres. *AI-Based Affective Music Generation Systems: A Review of Methods and Challenges*. *ACM Computing Surveys* 56(11), Article 287 (November 2024), 34 pages. <https://doi.org/10.1145/3672554>

3.6 Long-Term Cognitive Augmentation

Shoya Ishimaru (Osaka Metropolitan University, JP)

License  Creative Commons BY 4.0 International license
© Shoya Ishimaru

Human cognitive capabilities, such as learning, reasoning, and communicating, can be enhanced through various intervention approaches. While direct physiological interventions, including biofeedback and Transcranial Magnetic Stimulation (TMS) provide immediate cognitive enhancement during active stimulation, their effects might be typically transient and limited to the intervention period. My research proposes an alternative paradigm: achieving sustained cognitive augmentation by developing users’ meta-skills through information systems. Our approach leverages sensing technologies, including eye-tracking, physiological sensors, and behavioral analytics to recognize cognitive states during learning activities. By providing real-time feedback on engagement levels, comprehension, self-confidence, and

mental workload, these systems enable users to develop metacognitive awareness, the ability to monitor and regulate their own cognitive processes. Systems such as intelligent e-books that adapt content based on reader interest [1], engagement gauges that visualize attention in online meetings [2], and gaze-based word-count trackers for reading with smart eyewear exemplify this approach. Unlike temporary enhancement through direct neural stimulation, cultivating meta-skills creates lasting improvements in cognitive capability, as users internalize strategies for self-regulated learning, attention management, and effective knowledge acquisition that persist beyond system use.

References

- 1 R. Morita, K. Watanabe, J. Zhou, A. Dengel, and S. Ishimaru. *GenAIRreading: Augmenting Human Cognition with Interactive Digital Textbooks Using Large Language Models and Image Generation Models*. In *Proceedings of the Augmented Human International Conference 2025 (AHs '25)*, 2025.
- 2 K. Watanabe, T. Sathyanarayana, A. Dengel, and S. Ishimaru. *EnGauge: Engagement Gauge of Meeting Participants Estimated by Facial Expression and Deep Neural Network*. *IEEE Access*, pp. 52886–52898, 2023.
- 3 S. Ishimaru, K. Kunze, K. Kise, and A. Dengel. *The Wordometer 2.0: Estimating the Number of Words You Read in Real Life using Commercial EOG Glasses*. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '16 Adjunct)*, pp. 293–296, 2016.

3.7 Open-source neural wearables for the future of work

Michael Knierim (KIT – Karlsruher Institut für Technologie, DE)

License  Creative Commons BY 4.0 International license
© Michael Knierim

My research focuses on developing human-centered adaptive technologies through innovative neural wearables and large-scale biodata to sustainably enhance productivity, health, and well-being. A strong focus of my work is the development of discreet wearable systems (such as ear-EEG integrated headphones) and scalable software solutions that enable the transition from controlled experiments to everyday environments – solutions that I contribute as open-source resources to the scientific community. Working at the intersection of computer science, information systems, and neuroscience, I bridge wearable technology development with rigorous empirical research. My interdisciplinary approach addresses the fundamental challenge of creating reliable biosensing systems while ensuring these adaptive technologies genuinely serve human needs and well-being. Correspondingly the major theme in my work is understanding human cognitive and affective states through multimodal biosensing in both laboratory and real-world contexts. Through studies combining brain activity (EEG), eye tracking (EOG), heart rate variability (ECG/PPG), and behavioral data, I investigate states like cognitive load, flow, mental overload, and fatigue across settings ranging from controlled laboratory experiments to extended field studies in real-world knowledge work and home office scenarios.

At the Dagstuhl Seminar, I contributed a demonstration of the latest versions and ideas for wearable EEG headphones and explored potentials for novel use case scenarios, technical improvements and areas for collaboration like the use of wearable sensors in human-AI interaction or sensing context data to increase the reliability of everyday neural recordings.

References

- 1 T. King, M. Knierim, P. Lepold, C. Clarke, H. Gellersen, M. Beigl, and T. Röddiger. *earEOG via periauricular electrodes to facilitate eye tracking in a natural headphone form factor*. *Scientific Reports* 15(1), 2025. <https://doi.org/10.1038/s41598-025-16839-z>
- 2 M. T. Knierim, C. Zimny, G. Ivucic, and T. Röddiger. *Advancing wearable BCI: Headphone EEG for cognitive load detection in lab and field*. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)* 9(1), 2025. <https://doi.org/10.1145/3712283>
- 3 M. T. Knierim, F. Stano, F. Kurz, A. Heusch, and M. L. Wilson. *Exploring Flow in Real-World Knowledge Work Using Discrete cEEGrid Sensors*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, 2025. <https://doi.org/10.1145/3706598.3713512>

3.8 Estimating Human’s Cognitive Situation using AI and CV

Hideki Koike (*Institute of Science Tokyo, JP*)

License  Creative Commons BY 4.0 International license
© Hideki Koike

Cognitive sensing is sometimes very difficult and not reliable. In order to obtain reliable data, it is necessary to use relatively big sensors such as fMRI. On the other hand, recent advances in computer vision (CV) technology enables us to estimate human activities by using a small number of cameras. We are currently developing methods of estimating human muscle activities by using images captured by one camera. As a future vision, it might be possible to estimate a human’s internal state, not only the muscle activities but also brain activities by monitoring the face, eye gaze, and so on using the power of CV. Also the estimation would be more reliable by integrating with other sensor information such as heart rate, blood pressure, or else, which can be obtained by a small wearable device.

References

- 1 Ruofan Liu, Yichen Peng, Takanori Oku, Erwin Wu, Shinichi Furuya, and Hideki Koike. *PianoKeystroke-EMG: Piano Hand Muscle Electromyography Estimation from Easily Accessible Piano Keystroke*. In *SIGGRAPH Asia 2024 Posters (SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 6, pp. 1–2, 2024. <https://doi.org/10.1145/3681756.3697878>
- 2 Ruofan Liu, Yichen Peng, Takanori Oku, Chen-Chieh Liao, Erwin Wu, Shinichi Furuya, and Hideki Koike. *PiaMuscle: Improving Piano Skill Acquisition by Cost-effectively Estimating and Visualizing Activities of Miniature Hand Muscles*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 364, pp. 1–16, 2025. <https://doi.org/10.1145/3706598.3713465>
- 3 Ruofan Liu, et al. *From Pose to Muscle: Multimodal Learning for Piano Hand Muscle Electromyography*. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, to appear, 2025.

3.9 Artificial Intelligence for Amplifying Human Perception

Thomas Kosch (HU Berlin, DE)

License  Creative Commons BY 4.0 International license
© Thomas Kosch

My research explores how artificial intelligence can become an active collaborator in human perception and cognition. I study how computational models of behavior, physiology, and context can foster mutual understanding between humans and AI systems. By integrating signals from the body, such as interoceptive and proprioceptive cues, into interactive systems, I design and build interfaces that amplify human perception rather than replace it. A central theme in my work is amplified perception, the idea that AI can enhance our awareness of internal and environmental states, enabling more transparent, adaptive, and trustworthy interactions. This approach challenges the notion of AI as a controlling agent and instead positions it as a perceptual partner that co-regulates with the user. Through experimental and prototyping work in extended reality and wearable computing, I explore how feedback loops between humans and AI can enhance decision-making, empathy, and self-regulation. Another theme is understanding how the interaction with AI shapes the mental models and metacognition of users. For example, priming users about an AI augmentation system influences the user's view of their own performance, even when the AI system is non-functional. I refer to this as the placebo effect of AI, which should be carefully investigated when evaluating and deploying interactive AI systems, particularly in terms of user agency.

References

- 1 E. Bosch, D. Bethge, M. Klosterkamp, and T. Kosch. *Empathic technologies shaping innovative interaction: Future directions of affective computing*. In *Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference*, pp. 1–3, October 2022.
- 2 S. Villa, F. J. E. Krammer, Y. Weiss, R. Welsch, and T. Kosch. *Understanding the Influence of Electrical Muscle Stimulation on Motor Learning: Enhancing Motor Learning or Disrupting Natural Progression?*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, April 2025.
- 3 T. Kosch, R. Welsch, L. Chuang, and A. Schmidt. *The placebo effect of artificial intelligence in human–computer interaction*. *ACM Transactions on Computer-Human Interaction* 29(6), 1–32, 2023.
- 4 S. Villa, T. Kosch, F. Grelka, A. Schmidt, and R. Welsch. *The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior*. *Computers in Human Behavior* 146, 107787, 2023.
- 5 S. Villa, L. L. Barth, F. Chiossi, R. Welsch, and T. Kosch. *Whose mind is it anyway? A systematic review and exploration on agency in cognitive augmentation*. *Computers in Human Behavior: Artificial Humans*, Article 100158, 2025.
- 6 Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. *AI makes you smarter but none the wiser: The disconnect between performance and metacognition*. *Computers in Human Behavior* 175, 108779, 2026. <https://doi.org/10.1016/j.chb.2025.108779>

3.10 Reflexive Wearables: Turning Physiological Sensing into Human Insight

Kai Kunze (*Keio University – Yokohama, JP*)

License  Creative Commons BY 4.0 International license
© Kai Kunze

My research explores how physiological sensing can move from passive observation to active participation in human cognition and emotion. I focus on signals from the eyes, skin, and cardiovascular system, studying how they can reveal changes in attention, fatigue, or affective state. Through wearable platforms such as smart eyewear [1], I aim to integrate these signals into feedback systems that help people better understand and regulate their mental and emotional processes.

A major theme in my work is digitalized emotion, the idea that feelings and physiological reactions can be captured, represented, and re-experienced in meaningful ways. This approach opens possibilities for sharing emotional experiences, enhancing empathy, and supporting well-being. At the same time, I explore cognitive augmentation by developing tools that extend perception and awareness through subtle feedback rather than direct control. Working at the intersection of computer science, design, and the arts, I collaborate with engineers, performers, and psychologists to study how technology can act as both mirror and partner. My goal is to establish new forms of human-machine interaction that are reflective, adaptive, and emotionally aware, helping people not only to perform better but also to understand themselves and others more deeply.

References

- 1 Qing Zhang, Junyu Chen, Yifei Huang, Jing Huang, Thad Starner, Kai Kunze, and Jun Rekimoto. *Beyond Symbols: Motion Perception Cues Enhance Dual-Task Performance with Wearable Directional Guidance*. In *Proceedings of the 2025 ACM International Symposium on Wearable Computers (ISWC '25)*. Association for Computing Machinery, New York, NY, USA, 2025, pp. 22–29. <https://doi.org/10.1145/3715071.3750418>

3.11 Inferring Cognition through Oculo-metrics: From Laboratory Research to Real-World Applications

Diako Mardanbegi (*American University of Beirut, LB*)

License  Creative Commons BY 4.0 International license
© Diako Mardanbegi

Over the past 15 years, my research has focused on how eye movement metrics can be sensed and used for interaction with wearable systems [1, 2]. Decades of work have shown strong links between ocular metrics and cognitive states such as attention, fatigue, mental workload, stress, and focus, as well as neurological conditions. My current interest lies in understanding how these eye movement dynamics can support cognitive augmentation in future wearable technologies.

With the rise of eye-worn XR devices enabling continuous eye tracking, a key challenge is translating controlled lab findings into real-world settings. This issue extends beyond eye tracking: other neuro-sensing modalities like EEG, fNIRS, and EMG face similar gaps. Despite the growing availability of consumer brain-sensing devices (e.g., Mendi, MindWave), the mapping between neural signals and higher-order cognitive states such as focus or workload remains weakly supported. Direct, universal links between these signals and complex cognitive functions are still lacking.

At the Dagstuhl Seminar, we discussed cognitive sensing and augmentation from multiple perspectives, including ethics, diversity and inclusion, and technical challenges. One of the most thought-provoking discussions centered on the fundamental gap between sensing and interpretation. Specifically, cognitive states cannot be inferred in a straightforward, one-to-one manner from individual signals (e.g., “signal X implies cognitive state Y”), as these states are influenced by multiple contextual and individual factors, particularly in real-world environments.

Looking ahead, I envision a more reliable and ecologically valid approach that integrates multiple sources of low-level physiological and neural evidence from eye trackers, brain-sensing devices, and fitness trackers, with contextual information and user feedback. By correlating these multimodal signals with higher-level cognitive states and moods in a holistic, personalized framework, we can better capture individual variability, habits, and subjective experiences. Artificial intelligence will play a critical role in linking these multimodal data streams to cognitive biomarkers and potential health indicators, ultimately enabling more robust and meaningful cognitive augmentation through wearable technologies.

References

- 1 L. Sidenmark, D. Mardanbegi, A. R. Gomez, C. Clarke, and H. Gellersen. *Bimodalgaze: Seamlessly refined pointing with gaze and filtered gestural head movement*. In *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–9, June 2020.
- 2 D. Mardanbegi, B. Mayer, K. Pfeuffer, S. Jalaliniya, H. Gellersen, and A. Perzl. *Eye-seethrough: Unifying tool selection and application in virtual environments*. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 474–483. IEEE, March 2019.

3.12 Assistive Augmentation: Fundamentally Transforming Human Ability

Suranga Nanayakkara (National University of Singapore, SG)

License © Creative Commons BY 4.0 International license
© Suranga Nanayakkara

The rapid development of smart devices and emerging AI capabilities holds immense potential to transform not only our daily interactions but also our pursuit of life goals. Existing user interfaces, however, often fall short of realizing this potential. Individuals with disabilities, who rely heavily on assistive technologies for daily functioning and independence, face significant challenges due to the limitations of current approaches. Consider a person with visual impairment navigating a supermarket. While traditional accessibility features, such as text-to-speech, offer some assistance with shopping, they often require a cumbersome series of interactions, hindering the user’s ability to efficiently gather information and make choices. We argue that the perceived separateness and disconnect between accessibility devices and users hinder the experience of true empowerment.

Why is that? I believe that conversations have been overly focused on user-system coexistence and compatibility (e.g., “I will use this device” and “This is a user-friendly device”) rather than integration and augmentation of ability (e.g., “This device is a part of me” and “I have evolved with technology to unlock my full potential”). A paradigm shift is needed from viewing technology as a separate tool to considering it as an integrated extension of the human body, mind, and identity. We call this new perspective of human augmentation is called assistive augmentation [1].

Creating such Assistive Augmentations pose a threefold challenge as they require 1) exploring ways to seamlessly integrate into users body or behaviour, 2) deeper understanding of the user so as to provide the right level of support while keeping the user in control 3) holistic design approach to create meaningful applications. Inspired by insights from psychology, human centred machine learning, and design thinking, we have an unprecedented opportunity to design a new generation of Assistive Augmentations.

To highlight this, since 2011, I have been exploring several example project within this theme at the “Augmented Human Lab”. This include projects such as “AiSee” [2] that allows blind users to access information simply by pointing at objects and asking questions; “Prospero” [3] a memory training assistant that is able to detect when a user can learn more efficiently; “MuSS-Bits” [4], that allows deaf users to feel the music.

References

- 1 F. F. Tan, C. Gupta, D. P. D. Rajendran, P. Maes, and S. C. Nanayakkara. *Assistive Augmentation: Fundamentally Transforming Human Ability*. *Interactions* 32(1), 22–27, 2025. <https://doi.org/10.1145/3702656>
- 2 *Support Visually Impaired Grocery Shoppers*. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4(4), 1–25.
- 3 S. W. T. Chan, H. Zhang, and S. C. Nanayakkara. *Prospero: A Personal Wearable Memory Coach*. In *Proceedings of the 10th Augmented Human International Conference 2019*, pp. 1–5, March 2019.
- 4 B. Petry, T. Illandara, and S. C. Nanayakkara. *MuSS-bits: sensor-display blocks for deaf people to explore musical sounds*. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, pp. 72–80, November 2016.

3.13 Understanding Cognitive Augmentation through the Lenses of Experience, Agency, and Ethics

Jasmin Niess (University of Oslo, NO)

License  Creative Commons BY 4.0 International license
© Jasmin Niess

In a world where technology increasingly shapes how we think, feel, and connect, my research asks how we can design interactive systems that genuinely support people’s long-term well-being. One key area of my work focuses on cognitive augmentation such as neurotechnologies. I see this as a space for reimagining how humans and technologies can evolve together in ways that extend and enrich our cognitive and emotional capacities over time.

Within this space, my work addresses three interrelated challenges. The first concerns (i) bridging the gap between cognitive sensing data and people’s lived experience, for instance, by understanding when and how feedback should be provided and how individuals can create meaningful experiences with and from their data. The second explores (ii) agency and relationship building, examining how people experience agency when augmented and how they perceive such technologies, whether as tools, partners, or even as parts of themselves. The third centres on (iii) diversity, ethics, and regulation, highlighting the need to move beyond the narrow, Western-centric, neurotypical user profiles that dominate much of the current research landscape.

Drawing on my background in HCI and psychology, I develop theories, new measures, and interactive prototypes grounded in psychological knowledge. My work aims to translate conceptual insights into actionable guidance for researchers and practitioners, advancing a vision of technology that supports long-term well-being and meaningful augmentation experiences.

References

- 1 M. Bentvelzen, J. Niess, M. P. Woźniak, & P. W. Woźniak. *The development and validation of the technology-supported reflection inventory*. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 2021. <https://doi.org/10.1145/3411764.3445673>
- 2 S. Hoddø Bakås, M. P. Woźniak, J. Herstad, P. W. Woźniak, & J. Niess. *The Dual Model for Everyday Stress Technology: Understanding the Lived Experience of Data-Driven Stress*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, 2025. <https://doi.org/10.1145/3706598.3713174>
- 3 J. Niess & P. W. Woźniak. *Embracing Companion Technologies*. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, 2020. <https://doi.org/10.1145/3419249.3420134>
- 4 S. Villa, J. Niess, T. Nakao, J. Lazar, A. Schmidt, & T. Machulla. *Understanding Perception of Human Augmentation: A Mixed-Method Study*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 2023. <https://doi.org/10.1145/3544548.3581485>
- 5 N. Wagener, Y. Rogers, & J. Niess. *The Self-Care Technology Process Model (SCTpm): A Framework for Designing VR-Based Self-Care Technology*. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. ACM, 2025. <https://doi.org/10.1145/3715336.3735743>

3.14 AI-Powered Human Augmentation in the Workplace

Evangelos Niforatos (TU Delft, NL)

License © Creative Commons BY 4.0 International license
© Evangelos Niforatos

Evolution has always been the main driving force behind slow but steady change for the human brain and body. However, living in the Information era, our perceptual and cognitive capacities cannot simply rely on natural evolution to keep up with the immense advancements in emerging technologies. Tools we use daily (AI agents, Chatbots, PCs, smartphones, wearables, etc.) remain largely uninformed about our perceptual levels and cognitive states, forming a so-called “cognitive gap” between the human (user) and the machine (system) [1].

My research explores how AI-powered systems can sense and enhance human cognition in real-world productivity contexts, in an attempt to bridge the “cognitive gap” by utilizing physiological sensing, machine intelligence, and human-computer interaction principles. By combining sensing modalities such as EEG [2], eye-tracking [3], and physiological signals with generative and reasoning models, my work aims to interpret and respond to human cognitive processes, such as attention, recall, problem-solving, or creativity in situ. This approach contributes to developing adaptive, cognition-aware interfaces that support decision-making, learning, and productivity without compromising autonomy or well-being.

In the “Cognitive Sensing and Interaction” seminar theme, my interest lies in connecting cognitive sensing with generative AI (GenAI) to create systems that not only infer mental states but also proactively scaffold human cognition. For instance, integrating GenAI with

sensed cognitive workload can enable systems that offer context-aware explanations, personalized feedback, or cognitive offloading mechanisms in manufacturing [4], healthcare, and education. Ultimately, this research seeks to advance human-centered cognitive augmentation by integrating intelligent systems into workflows, where AI complements rather than replaces human intelligence.

References

- 1 E. Niforatos, A. Vourvopoulos, and M. Langheinrich. *Amplifying human cognition: bridging the cognitive gap between human and machine*. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pp. 673–680, September 2017.
- 2 E. Niforatos, T. He, A. Vourvopoulos, and M. Giannakos. *Democratizing EEG: Embedding electroencephalography in a head-mounted display for ubiquitous brain-computer interfacing*. *International Journal of Human-Computer Interaction* 41(11), 7015–7039, 2025.
- 3 K. Sharma, E. Niforatos, M. Giannakos, and V. Kostakos. *Assessing cognitive performance using physiological and facial features: Generalizing across contexts*. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4(3), 1–41, 2020.
- 4 S. Kernan Freire, M. Foosherian, C. Wang, and E. Niforatos. *Harnessing large language models for cognitive assistants in factories*. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pp. 1–6, July 2023.

3.15 From Perception to Shared Cognition: Leveraging Egocentric Physiological and Behavioral Cues for Empathic Remote Collaboration

Prasanth Sasikumar (National University of Singapore, SG)

License  Creative Commons BY 4.0 International license
© Prasanth Sasikumar

Remote collaboration in mixed reality often lacks the rich, implicit signals that ground face-to-face interaction, particularly those tied to cognitive and affective states. My research bridges this gap by integrating real-time physiological sensing with egocentric perceptual data (e.g., eye gaze, hand gestures, and first-person scene understanding) to reconstruct not only the physical workspace but also the cognitive context of collaborators. By fusing signals such as electrodermal activity (cognitive load), heart-rate variability (arousal), and EEG-derived attention with behavioral cues captured from head-mounted sensors, I design systems that support mutual awareness, adaptive communication, and empathic co-experience.

This approach builds on empirical findings showing that sharing combined physiological cues significantly enhances co-presence and social awareness in MR collaboration [1]. Critically, I treat these signals not as isolated metrics, but as interdependent markers of situated cognition: for instance, aligning gaze fixations with spikes in cognitive load can reveal moments of confusion or insight, while synchronizing gesture fluency with attention metrics can indicate confidence or hesitation. Such multimodal inference enables systems that scaffold not just task performance, but shared understanding, turning remote collaboration into a co-regulated cognitive process.

Within the context of the Dagstuhl Seminar on cognition and interaction, I am particularly interested in advancing frameworks that tightly couple perception (what is sensed from the user’s egocentric viewpoint) with cognition (what those signals imply about mental state).

My goal is to contribute to a vision of human–machine interaction where technology helps collaborators think and feel together, using egocentric sensing as a foundation for reflective, adaptive, and cognitively attuned remote experiences.

References

- 1 P. Sasikumar et al. *A User Study on Mixed Reality Remote Collaboration with Physiological Sensing*. In *IEEE VR 2023*, 2023.
- 2 R. Welsch et al. *AI makes you smarter but none the wiser: The disconnect between performance and metacognition*. *Computers in Human Behavior*, 2025.
- 3 S. Ishimaru. *Long-Term Cognitive Augmentation*. Dagstuhl Seminar Abstract.
- 4 K. Kunze. *Reflexive Wearables: Turning Physiological Sensing into Human Insight*. Dagstuhl Seminar Abstract.

3.16 Designing Cognition-Aware Interfaces for Neurotechnology

Christina Schneegass (TU Delft, NL)

License © Creative Commons BY 4.0 International license
© Christina Schneegass

People are increasingly eager to know more about themselves through technology, which has so far focused on physiological data tracking. Advances in wearable technology and AI are enabling consumer neurotechnology to measure cognitive activity, presenting new research opportunities in Human-Computer Interaction (HCI). I am bringing together knowledge from cognitive psychology, human-centered design, and computer science to explore how we can design interfaces and interaction techniques for cognition-aware technologies – specifically non-invasive consumer neurotechnologies (CNTs). The goal is to help people understand and reflect on cognitive processes, such as focus, stress, and learning, to improve their mental health and well-being. I follow a human-centered approach, studying how people want to engage with cognitive data to design consumer neurotechnologies that respond to people’s needs and support healthy reflection.

The seminar topic cognitive sensing and interaction aligns with my research ambitions, especially considering how we can balance inaccuracies in sensing with the users need for explanations and accuracy, in particular when there is a mismatch between objectively sensed data and subjective experiences.

References

- 1 C. Schneegass, M. L. Wilson, J. Shaban, J. Niess, F. Chiossi, T. Mitrevska, and P. W. Woźniak. *Broadening the mind: how emerging neurotechnology is reshaping HCI and interactive system design*. *i-com* 23(2), 165–177, 2024.
- 2 T. Mitrevska, C. Kobiella, J. Feckl, S. Sakel, A. Butz, and C. Schneegass. *Investigating the effects of neurofeedback on knowledge workers’ perceptions of self-accomplishment, focus, and task performance*. In *Proceedings of the 37th Australian Conference on Human-Computer Interaction (OZCHI ’25)*, November 29–December 03, 2025, Sydney, Australia. ACM, New York, NY, USA, 13 pages, 2025. <https://doi.org/10.1145/3764687.3764696>
- 3 Y. Chen, T. Mitrevska, Tilman Dingler, and C. Schneegass. *Understanding Users’ Perceptions and Barriers to Mental Workload Self-Tracking*. In *Proceedings of the 37th Australian Conference on Human-Computer Interaction (OZCHI ’25)*, November 29–December 03, 2025, Sydney, Australia. ACM, New York, NY, USA, 8 pages, 2025. <https://doi.org/10.1145/3764687.3769928>

3.17 Neuroadaptive and Embodied Interfaces for Cognitive Ergonomics

Benjamin Tag (UNSW – Sydney, AU)

License  Creative Commons BY 4.0 International license
© Benjamin Tag

My research investigates how technology can extend and refine human cognition through adaptive, embodied interaction. I focus on how mixed and virtual reality systems, combined with physiological sensing, can measure and respond to mental states such as fatigue, attention, and cognitive load. By integrating data from the eyes, muscles, and autonomic nervous system, I aim to build immersive environments that dynamically adjust to users' capabilities, augmenting their awareness, resilience, and performance rather than merely tracking behaviour. A central theme in my work is cognitive ergonomics: understanding how digital systems can support, rather than strain, the human mind. I explore how feedback mechanisms, visual, haptic, or neuroadaptive, can promote sustainable engagement and recovery in cognitively demanding contexts such as training, remote operations, and human-AI collaboration. These systems blur the boundary between observation and augmentation, treating physiological signals as active interfaces that mediate understanding between body and machine. Working across computer science, psychology, and design, I collaborate with neuroscientists, optometrists, and engineers to develop technologies that act not only as tools but as cognitive partners. My goal is to establish a foundation for ergonomic cognitive augmentation: technologies that help humans think, perceive, and adapt more effectively, while preserving well-being and agency in the face of increasing digital complexity.

References

- 1 Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. *A Critique of Electrodermal Activity Practices at CHI*. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 177, pp. 1–14, 2021. <https://doi.org/10.1145/3411764.3445370>
- 2 Ebrahim Babaei, Tilman Dingler, Benjamin Tag, and Eduardo Velloso. *Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around Mental Workload Measurement*. *International Journal of Human-Computer Studies* 201, 2025. <https://doi.org/10.1016/j.ijhcs.2025.103515>
- 3 Benjamin Tag, Andrew W. Vargo, Aman Gupta, George Chernyshov, Kai Kunze, and Tilman Dingler. *Continuous Alertness Assessments: Using EOG Glasses to Unobtrusively Monitor Fatigue Levels In-The-Wild*. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Paper 464, pp. 1–12, 2019. <https://doi.org/10.1145/3290605.3300694>
- 4 Benjamin Tag, Zhanna Sarsenbayeva, Anna L. Cox, Greg Wadley, Jorge Goncalves, and Vassilis Kostakos. *Emotion trajectories in smartphone use: Towards recognizing emotion regulation in-the-wild*. *International Journal of Human-Computer Studies* 166, 102872, 2022. <https://doi.org/10.1016/j.ijhcs.2022.102872>

3.18 Beyond Enhancement: Cognitive Sensing Futures

Steeven Villa (LMU München, DE)

License © Creative Commons BY 4.0 International license
© Steeven Villa

Augmentation technologies would likely increase adoption in the coming decades, thus understanding their behavioral and societal impacts is critical. My research examines how augmentation technologies influence human behavior and societal dynamics through a prototypical approach [1, 2]: implementing systems, analyzing their effects on human behavior [3] and social contexts [4], and synthesizing insights about short- and long-term impacts. My work has revealed that high expectations of improvement using such technologies increase risk-taking behavior and impact how people physiologically integrate information [5]. Interestingly, my investigation of placebo effects has shown that these positive responses are resistant to negative narratives, meaning a placebo effect could not be induced in augmentation technology contexts [6]. This work establishes a foundation for understanding both individual psychological responses and broader societal implications of human augmentation.

The Dagstuhl Seminar on cognitive sensing brought together researchers to address future challenges in this emerging field. I hosted a session exploring the trajectory of cognitive sensing technologies across three timeframes: short (5 years), medium (10 years) and long-term (15+ years). Through structured discussions with seminar attendees, we identified surveillance and potential abuse by corporate or institutional stakeholders as the most critical concerns requiring immediate attention. Specific scenarios that emerged included workplace surveillance and the abuse of physiological data capture to infer people's interests, attention, and cognitive states without consent. The session highlighted that technical advancement must be accompanied by robust alternatives against misuse to ensure responsible development of cognitive sensing technologies.

The insights from the seminar revealed gaps in understanding how unequal access to augmentation technologies shapes society. My next research phase will focus on two interconnected areas: first, analyzing asymmetric interactions between individuals with and without access to augmentation technologies to map societal impacts; and second, developing frameworks for establishing and maintaining trust in augmentation technologies. This work will provide both empirical evidence of technology-driven inequality and practical guidelines for equitable deployment of human augmentation systems.

References

- 1 Steeven Villa, Yannick Weiss, Mei Yi Lu, Moritz Ziarko, Albrecht Schmidt, and Jasmin Niess. *Envisioning Futures: How the Modality of AI Recommendations Impacts Conversation Flow in AR-enhanced Dialogue*. In *Proceedings of the 26th International Conference on Multimodal Interaction*.
- 2 Steeven Villa, Finn Jacob Eliyah Krammer, Yannick Weiss, Robin Welsch, and Thomas Kosch. *Understanding the Influence of Electrical Muscle Stimulation on Motor Learning: Enhancing Motor Learning or Disrupting Natural Progression?*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.
- 3 Chenxinran Shen, Jurgis Karpus, Thomas Kosch, Daniela Fernandes, Beatriz Mello, Robin Welsch, and Steeven Villa. *The Impact of Asymmetric AI Assistance on Decision-Making in Social Dilemmas: A Study on Human Augmentation in Economic Games*. In *Proceedings of the Augmented Humans International Conference 2025*, 2025.
- 4 Steeven Villa, Jasmin Niess, Takuro Nakao, Jonathan Lazar, Albrecht Schmidt, and Tonja-Katrin Machulla. *Understanding perception of human augmentation: A mixed-method study*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

- 5 Steeven Villa, Thomas Kosch, Felix Grelka, Albrecht Schmidt, and Robin Welsch. *The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior*. *Computers in Human Behavior*, 2023.
- 6 Agnes M. Kloft, Robin Welsch, Thomas Kosch, and Steeven Villa. “AI enhances our performance, I have no doubt this one will do the same”: The Placebo effect is robust to negative descriptions of AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

3.19 AI-Augmented Cognition and Co-Experience in Extended Reality

Robin Welsch (Aalto University, FI)

License  Creative Commons BY 4.0 International license
© Robin Welsch

My research examines how multimodal sensing and artificial intelligence can make people smarter. I want to combine psychophysiological and behavioral signals, eye movements, electrodermal activity, heart-rate variability, EEG, motion, voice, and context, to infer metacognitive states to build technologies that can support users in self-regulating human-AI interaction. Our recent research could show that metacognition is the bottle neck for performance in human-AI interaction [1] and we know that we can improve performance by adaptive UI by means of physiological sensing [2]. Therefore, I believe that the next step is to build metacognitively-adaptive interfaces to AI.

A second focus is measuring and shaping co-experience in digital environments [3], particularly social XR. I operationalize co-experience as the dynamic coupling of cognitive and affective states across people, quantified through physiological synchrony, and validated reports of presence and co-presence. At Dagstuhl, we have discussed, especially with Kai Kunze, which different analysis strategies and sensors could be used to quantify co-experience.

References

- 1 Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. *AI makes you smarter but none the wiser: The disconnect between performance and metacognition*. *Computers in Human Behavior*, 2025, Article 108779.
- 2 Francesco Chioffi, Yagiz Turgut, Robin Welsch, and Sven Mayer. *Adapting Visual Complexity Based on Electrodermal Activity Improves Working Memory Performance in Virtual Reality*. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 196 (September 2023), 26 pages. <https://doi.org/10.1145/3604243>
- 3 Agnes Mercedes Kloft, Julian Marcel Fabinc, Linda Hirsch, et al. *It’s Always Better When We’re Together: Physiological Synchrony and Shared Experiences in Social VR*. Preprint, 12 October 2025. <https://doi.org/10.21203/rs.3.rs-7134151/v1>

3.20 The Future of Cognitive Personal Informatics

Max L. Wilson (University of Nottingham, GB)

License © Creative Commons BY 4.0 International license
© Max L. Wilson

My view of consumer neurotechnology is that all of the current innovation is lab experiments and machine learning of datasets for accuracy etc are all a small step in a bigger view of our future living with consumer neurotechnology. This future is one where we have as many apps and wearables for our brain as we do for fitness programmes and long-term health tracking. Physical activity tracking has studied this long-term real-world use of wearable tech in the field of Personal Informatics. I have co-led the community of people that now look at Cognitive Personal Informatics [2, 4, 5] and the impact this will have on the academic landscape [3].

As part of this vision, we began work in 2018, performing longitudinal fieldwork with people tracking their own Mental Workload (as our initial theoretical framework). Initially this was self-reported through an app, and led to our introduction of the Mental Workload Cycle [1] – as a model for people should reach optimal mental effort levels across a whole day. This is distinct from most physical activity tracking that assumes more is always better, but closer to now more expert models of training support that models both activity and recovery.

Our in-progress work now involves longitudinal fieldwork with people that do have viable consumer neurotech. Over the last 12 months, we have engaged in a year of public engagement on consumer neurotechnology. Over 13 events, we have given 5,000 members of the public, teachers, children, parents, lawyers, policy advisors, and politicians the chance to try neurotechnology, and spoken to at least 5,000 more. My research vision is to move first-generation consumer neurotechnology towards devices that help people set goals for the brain, cognitive, and mental health, and can meaningfully communicate progress towards them.

References

- 1 S. Midha, M. L. Wilson, and S. Sharples. *Lived experiences of mental workload in everyday life*. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, April 2022.
- 2 C. Schneegass, M. L. Wilson, H. A. Maior, F. Chiossi, A. L. Cox, and J. Wiese. *The Future of Cognitive Personal Informatics*. In *Proceedings of the 25th International Conference on Mobile Human-Computer Interaction*, pp. 1–5, September 2023.
- 3 C. Schneegass, M. L. Wilson, J. Shaban, J. Niess, F. Chiossi, T. Mitrevska, and P. W. Woźniak. *Broadening the mind: how emerging neurotechnology is reshaping HCI and interactive system design*. *i-com* 23(2), 165–177, 2024.
- 4 M. L. Wilson, S. Midha, H. A. Maior, A. L. Cox, L. L. Chuang, and L. D. Urquhart. *SIG: moving from brain-computer interfaces to personal cognitive informatics*. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–4, April 2022.
- 5 M. L. Wilson, J. Shaban, H. A. Maior, C. Schneegass, and A. L. Cox. *The CHI'24 Workshop on the Future of Cognitive Personal Informatics*. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–6, May 2024.

3.21 Co-designing and Personal Fabrication of Assistive Technologies

Anusha Withana (*The University of Sydney, AU*)

License  Creative Commons BY 4.0 International license
© Anusha Withana

My research focuses on developing personalized wearable technologies that assist individuals with disabilities and chronic health conditions through sensing, actuation, and collaborative physical interactions. A key aspect of this work is engaging end users in the design and fabrication process using co-design tools. This participatory approach ensures that the resulting technologies are not only functional but also meaningful and empowering for the people who use them.

A key aspect that was missing in my research was cognitive considerations of co-design, actuation, and collaborative physical interactions. Critical questions include: How can cognitive sensing enhance creative co-design processes? How do we manage the complexities and adaptability of physical wearable devices without cognitive overload? How can cognitive sensing provide better control and agency over wearables? My main goal in attending the seminar was to broaden my research scope to address these challenges and integrate cognitive dimensions into wearable design.

In the seminar, I gained insights into the state of the art in cognitive sensing, both in commercial applications and academic research. I connected with new colleagues for potential collaborations, particularly in embedding cognitive sensing into co-design processes. These discussions opened exciting opportunities to explore how cognitive sensing can enhance accessibility and physical embodiment, how current cognitive sensing technologies can be co-designed and made with personal fabrication approaches and future ethical challenges that are worth exploring in the area.

References

- 1 Hongyu Zhou, Tom Kip, Andrea Bianchi, Zhanna Sarsenbayeva, and Anusha Withana. *Juggling Extra Limbs: Identifying Control Strategies for Supernumerary Multi-Arms in Virtual Reality*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (Best Paper Hon. Mention) (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 2025.
- 2 Hongyu Zhou, Treshan Ayesh, Chenyu Fan, Zhanna Sarsenbayeva, and Anusha Withana. *CoplayingVR: Understanding User Experience in Shared Control in Virtual Reality*. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8(3), September 2024.
- 3 Jiakun Yu, Supun Kuruppu, Biyon Fernando, Praneeth Bimsara Perera, Yuta Sugiura, Sriram Subramanian, and Anusha Withana. *IrOnTex: Using Ironable 3D Printed Objects to Fabricate and Prototype Customizable Interactive Textiles*. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8(3), September 2024.
- 4 Stephen Lin, Nisal Menuka Gamage, Kithmini Herath, and Anusha Withana. *MyoSpring: 3D Printing Mechanomyographic Sensors for Subtle Finger Gesture Recognition*. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '22)*, pp. 1–11. ACM, New York, NY, USA, February 2022.
- 5 Anusha Withana, Daniel Groeger, and Jürgen Steimle. *Tacttoo: A Thin and Feel-Through Tattoo for On-Skin Tactile Output*. In *The 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*, pp. 365–378. ACM Press, New York, NY, USA, 2018.

4 Working groups

4.1 Envisioning Novel Cognitive Interaction Scenarios

Thomas Kosch (HU Berlin, DE)

License  Creative Commons BY 4.0 International license
© Thomas Kosch

The second session was organized as a hike to the remnants of Castle Dagstuhl, with the main goal of creating space for broader and more speculative thinking while still keeping the discussion grounded in cognitive sensing features. The hike was structured as a walk-and-talk activity, where pairs of participants discussed future interaction scenarios and used the movement and changing setting as a deliberate prompt to explore alternatives. We encouraged each pair to focus on a specific context of use, such as collaboration, learning, safety-critical work, or creative tasks, and then to ask what cognitive information could add value in that context and what the associated risks would be.

The discussions moved from the “happy path” to edge cases, for example, what happens when cognitive inferences are wrong, when the user does not want to be sensed, or when the system’s adaptation becomes distracting. People also compared scenarios across domains and noticed repeating design tensions, such as personalization versus generalizability, helpful adaptation versus loss of agency, and subtle feedback versus transparency. After returning from the hike, the pairs shared short summaries, which helped us collect a wider set of scenario ideas than we would likely have generated in a room-based format. The outcome was a richer map of possible applications, along with a more explicit understanding of what design constraints and ethical considerations must be addressed early.

4.2 Metrics and Best Practices for Cognitive Sensing

Thomas Kosch (HU Berlin, DE)

License  Creative Commons BY 4.0 International license
© Thomas Kosch

During the first session, challenges emerged in utilizing cognitive sensing in everyday contexts, and participants began to share their best practices for ensuring reliability and collecting the most accurate data. In this second session, we built on this discussion by examining measures, analysis approaches, and opportunities and pitfalls across both laboratory and real-world settings. The group discussions were guided by cognitive measures, such as attention, workload, and fatigue, and participants completed the next Miro section. Our later analysis revealed that groups agreed that cognitive data must be collected in a context-aware, multimodal manner and interpreted in relation to an individual’s ongoing activity. They emphasized that cognitive data alone is insufficient and must be put into the context of the task, environment, and user state, using other sensor data. Examples for best practices for successful everyday cognitive sensing included feature fusion, motion and noise artifact filtering, data labeling approaches, and multimodal data integration to improve interpretability. Creating individualized baseline measures was also noted as an important best practice, as well as the use of contextual cues. Lastly, participants discussed the opportunities and pitfalls associated with both laboratory and real-world deployments. They noted the benefit of laboratory settings, as they allow for controlled conditions, clear ground

truth, and easier sensor management. However, they state that in these conditions, the resulting models might fail if applied outside the laboratory task and setting. In contrast, collecting data in everyday contexts yields more complex observations and higher relevance, but may be challenging to interpret, have power constraints, and raise privacy implications. It is important to consider where the final technology is supposed to be deployed, what the user needs are, and what sensing capabilities are actually needed for it.

4.3 Prototyping Cognitive Interaction Scenarios

Kai Kunze (Keio University – Yokohama, JP) and Thad Starner (Georgia Institute of Technology – Atlanta, US)

License  Creative Commons BY 4.0 International license
© Kai Kunze and Thad Starner

The second day began by explicitly extending the third session from the first day, transitioning from a conceptual discussion to concrete interaction scenarios that could be implemented and experienced. Participants brought their own devices and were also provided with hardware on-site, allowing us to directly test ideas rather than just discuss them. We worked in small groups around existing prototypes and early mock-ups, using short iteration cycles. We defined a scenario, decided which cognitive sensing features were assumed to be available, implemented or simulated a minimal interaction, and then reflected on what worked and what broke down.

A recurring focus was on making cognitive sensing actionable in the interaction, for example, by determining what the system should do when it detects a change in attention, workload, or uncertainty, and how that information should be communicated to the user. This hands-on work also highlighted the limitations very clearly. The participants repeatedly encountered questions about signal quality, calibration time, latency, and variability across different people and contexts. By the end of the session, the scenarios were much more specific, including clearer assumptions about sensing reliability, explicit points where uncertainty needs to be communicated, and a better understanding of what kinds of prototypes are feasible to pursue in the short term. The visions were presented by the participants, including a scenario, utopian, and dystopian views.

4.4 Policies for Cognitive Sensing and Cognitive Interaction

Jasmin Niess (University of Oslo, NO) and Thomas Kosch (HU Berlin, DE)

License  Creative Commons BY 4.0 International license
© Jasmin Niess and Thomas Kosch

The third day was structured as a single, longer session, intentionally shifting the seminar focus from designing interaction scenarios to the broader question of how cognitive sensing and cognitive interaction technologies should be governed. We began by identifying and unpacking the technical challenges that directly affect both feasibility and risk. Participants discussed the robustness of models outside controlled lab settings, the need for calibration and personalization, sensor noise and missing data, and the fact that many cognitive signals are indirect proxies that can be confounded by context, individual differences, fatigue, stress, or emotion. This led to a discussion of how easily systems can overclaim what they “know”

about a user, and why uncertainty needs to be represented explicitly, not only in the model but also in the interaction. We also discussed technical design choices that have policy implications, such as on-device versus cloud processing, what is stored versus processed transiently, and how model updates or domain shifts can alter system behavior over time.

From there, the session expanded to societal and legal challenges, with a focus on what makes cognitive data and cognitive inferences particularly sensitive compared to more traditional forms of interaction data. We discussed risks related to autonomy and manipulation, particularly in settings characterized by power asymmetries, such as workplaces, education, healthcare, and public services. We discussed how the AI act can help with regulating in this context. Furthermore, participants raised questions around consent that is nominally “informed” but not meaningful in practice, especially when sensing is passive or when people cannot realistically opt out. We also reflected on how cognitive interaction techniques could amplify bias or unfairness, for example, when models are trained on non-representative populations or when certain users systematically produce different signals. Legal aspects came up in relation to privacy, data protection, and accountability, including the challenge of regulating not only raw sensor data but also derived inferences about mental states. Even when data is processed locally, the downstream effects of an inference can be high-stakes. Therefore, the group discussed what accountability should look like when harm arises from an adaptive system rather than a single explicit decision.

Based on these discussions, we transitioned into policy ideation with the explicit goal of promoting the benevolent use of cognitive interaction techniques. We brainstormed concrete policy directions, such as requiring clear disclosure when cognitive sensing is active, providing user-facing explanations of what is inferred and at what confidence level, and enabling meaningful user control, including the ability to pause sensing, review what has been collected or inferred, and delete data. Another cluster of ideas focused on minimizing harm by limiting data retention, restricting secondary use, and preventing function creep, for example, avoiding the reuse of cognitive inferences for performance evaluation, targeted persuasion, or surveillance. Participants also discussed the need for deployment constraints, such as stricter requirements for high-risk contexts, stronger oversight when there is an imbalance of power, and explicit rules for when cognitive adaptation should be disallowed or require independent review. Throughout, we have attempted to make these policies actionable by connecting them to implementation mechanisms, such as auditing, documentation requirements, standardized reporting of uncertainty and failure cases, and evaluation protocols that encompass user experience and societal impact, in addition to performance metrics.

In a dedicated part of the session, we then mapped policies to stakeholders and debated who should realistically implement and enforce them. We covered researchers and academic venues, for example, through ethics guidelines, artifact documentation standards, and expectations for responsible evaluation. We discussed how companies build sensing hardware and platforms, as many policy-relevant choices are embedded in default settings, APIs, and data flows. We also included deployers such as employers, schools, and healthcare institutions, who often define the actual usage conditions and incentives. Finally, we considered the role of standards bodies, regulators, and funding agencies in shaping baseline requirements, compliance pathways, and incentives for safer design.

The session concluded with participants writing brief abstracts that summarized their research background and perspectives that contributed to the Dagstuhl Seminar.

4.5 Feasibility of Cognitive Sensing Methods in Everyday Life

Christina Schneegass (TU Delft, NL)

License  Creative Commons BY 4.0 International license
© Christina Schneegass

On the first day, we focused on the practical applications of cognitive sensing technologies and asked our participants to share their experiences using them in personal/research settings. Our aim was to collect devices, services, and technologies that work well, as well as those that don't work at all, for everyday situations. We sorted the devices using a feasibility–usefulness matrix using a Miro Board. The participants collected a wide variety of devices and sensing modalities (e.g., wearable neurophysiology, behavioral proxies, environmental sensors). We observed that the experiences of participants, as well as their understanding of what constitutes “reliable,” varied among participants. It became clear that the feasibility of using a sensing device outside of laboratory settings depended heavily on the context, the data required, and the protocol. Our participants reported using various research and consumer devices in their practice in everyday contexts, including smartwatches and earbuds. Others, such as EEG and fNIRS, were rated as unreliable outside of laboratory conditions. We further identified challenges that make it difficult to utilize cognitive sensing technologies in the wild, including noise sensitivity, data ambiguity, interpretation challenges, and ethical and privacy concerns (e.g., devices being perceived as surveillance by users rather than as assistance). However, all participants agreed that cognitive sensing technologies must develop to a state where they can be reliably used in everyday contexts to collect valid and reliable data. In essence, this session provided an overview of our seminar participants' experiences with different devices, preferences, and challenges, which we explored in more depth later that day.

4.6 Vision for Cognitive Sensing in 5, 15, and 50 Years

Steeven Villa (LMU München, DE)

License  Creative Commons BY 4.0 International license
© Steeven Villa

Ultimately, our goal was to gather participants' visions for cognitive sensing technologies over the next 5, 15, and 50 years. We asked participants to sketch scenarios using a vignette style and AI-generated visualizations to support their idea. The groups focused on various areas, including ADHD-related support, cognitive monitoring for truck drivers, neural interfaces, emotional regulation tools, and minimally invasive brain-sensing technologies. The 5-year scenarios described ideas for everyday contexts based on research and devices that are (almost) realistic in laboratory settings today. For example, adapting audio environments to improve focus for individuals with ADHD, or detecting fatigue in driving contexts. The 15 and 50-year scenarios were increasingly speculative, embedding cognitive sensing into everyday context (e.g., ubiquitous monitoring, automated workload management), and combining it with neurostimulation for cognitive modulation. Through very precise portable sensing and neural intervention, our groups envisioned that we would be able to suppress disruptions, turn off parts of our brain, and have fully integrated cognition-adaptive personal systems.

4.7 Developing a Framework for Cognitive Interaction

Max L. Wilson (University of Nottingham, GB)

License © Creative Commons BY 4.0 International license
© Max L. Wilson

The third session shifted the focus from generating scenarios to establishing shared foundations by discussing what we actually mean by “cognition” in the context of sensing and interaction design. We spent a substantial amount of time aligning terminology, as participants came from different backgrounds and used overlapping terms in varying ways. We discussed how cognition relates to, but is not identical with, constructs such as attention, mental workload, intention, learning, fatigue, affect, and arousal, and we examined where it makes sense to draw boundaries for the purpose of designing interactive systems. This led to a discussion of levels of inference, ranging from low-level signals and proxy measures to higher-level cognitive states and predictions about behavior.

We also talked about the role of context and individual differences, and why a framework needs to represent uncertainty and avoid treating inferred cognition as a stable ground truth. From there, we began shaping the core elements of a framework for sensing cognition, including how sensing signals are mapped to cognitive constructs, how those constructs are represented and updated over time, and how they should influence interaction choices such as feedback, adaptation, and user control. We emphasized that evaluation needs to go beyond classification accuracy and should include user experience outcomes, such as comprehensibility, perceived control, trust, and the consequences of errors. By the end, we established a clearer shared definition to work with and an initial structure that can connect sensing capabilities to interaction mechanisms in a way that is explicit about cognitive assumptions, possibilities, and limitations.

5 Panel discussions

5.1 The Future of Wearable Neurotech: From Sensing to Neuromorphic Intelligence

Mustafa Hamada (Mendi Innovations – Stockholm, SE)

License © Creative Commons BY 4.0 International license
© Mustafa Hamada

This keynote explores the evolution of consumer neurotechnology from basic monitoring to intelligent, adaptive systems. The \$1.8 trillion wellness market is witnessing a paradigm shift toward wearables with clinical efficacy, AI-driven personalization, and continuous bio-monitoring. Current consumer neurotechnology products, such as Mendi, demonstrate how functional near-infrared spectroscopy (fNIRS) can make brain wellness accessible through gamified neurofeedback and cognitive tracking. Existing consumer neurotechnology products face multiple critical limitations, including constrained battery life, reliance on cloud infrastructure raising privacy concerns, and insufficient edge computing for real-time adaptation. The next frontier lies in neuromorphic computing architectures that mimic the brain’s own neural efficiency in signal processing. Drawing on principles from dendritic computation, neuromorphic processors consume less than 1% of the power of conventional AI systems while enabling sophisticated on-device intelligence. This enables truly wearable devices such

as NEXA Mind for seizure prediction and NEXA Heart for cardiac monitoring systems that operate continuously, privately, and independently of network connectivity. As neurotechnology transitions from clinical to consumer contexts, we must address fundamental questions about meaningful brain engagement, scalable low-power architectures, and safeguarding user agency and trust.

5.2 Our Future Living Scenarios with Personal Brain Scanners

Max L. Wilson (University of Nottingham, GB)

License  Creative Commons BY 4.0 International license
© Max L. Wilson

In this presentation, I will reflect on a year of public engagement, during which we gave 5,000 people the chance to experience consumer neurotechnology for the first time and spoke to a further 5,000 people at 13 free science festivals across the UK. In particular, I will unpack the implications raised for scenarios of use and the kinds of meaning people want to make from the data. I will combine these motivating scenarios that came from recent cognitive personal informatics workshops at SIGCHI conferences. The primary novel insights for consumer neurotechnology, especially with respect to non-medical scenarios, were related to use in schools, use by people with neurodiversity, and use by the elderly. Beyond ethical concerns of the use of neurotechnology to enforce learning in schools, there were two occasions where children were happier for neurotechnology to be used in schools: 1) in order for teachers to better understand how neurodivergent learners were behaving, and 2) as feedback to show teachers that their classes are boring. Both of these flip the value to the student's benefit, rather than helping teachers and parents to enforce learning. Neurodivergent people wanted consumer neurotech to help track and understand themselves better, especially to provide evidence that, e.g., “new meds are working”. Older users wanted a warning system: one that consumer neurotech would tell them that their brain data looks worrying, and it's time to see a doctor. Aside from these specific cases, common themes were: 1) that people wanted measurement over time, 2) to show progress towards a goal, and 3) that people wanted the associated apps to tell them if their brain data is good or bad. All three of these general points have strong implications for industry and researchers, where relevant goals, showing progress, and recognising good and bad data are all currently unknown.

6 Open problems

6.1 Day 1 Synopsis

Thomas Kosch (HU Berlin, DE), Kai Kunze (Keio University – Yokohama, JP), Christina Schneegass (TU Delft, NL), and Thad Starner (Georgia Institute of Technology – Atlanta, US)

License  Creative Commons BY 4.0 International license
© Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner

On the first day, we investigated the feasibility, challenges, and future research opportunities of cognitive sensing. Moreover, we investigated how sensing strategies can be translated into interaction concepts. The first session explored the feasibility of cognitive sensing, including various sensing modalities (e.g., eye tracking, electroencephalography, and electrodermal

activity), and discussed how they can be realistically integrated into everyday living scenarios. The second session investigated how useful features can be extracted from the discussed sensory measurements. The third session envisioned future interaction concepts.

6.2 Day 2 Synopsis

Thomas Kosch (HU Berlin, DE), Kai Kunze (Keio University – Yokohama, JP), Christina Schneegass (TU Delft, NL), and Thad Starner (Georgia Institute of Technology – Atlanta, US)

License © Creative Commons BY 4.0 International license
© Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner

The second day extended the third session from the first day by envisioning concrete interaction scenarios. Participants brought and were provided with hardware to test and extend existing prototypes. The second session included a hike to the remnants of Castle Dagstuhl. This hike consisted of a walk-and-talk session, where pairs of participants discussed future interaction scenarios based on cognitive sensing features. The third session focused on unifying a common understanding of cognitive interaction, resulting in a framework for the future.

6.3 Day 3 Synopsis

Thomas Kosch (HU Berlin, DE), Kai Kunze (Keio University – Yokohama, JP), Christina Schneegass (TU Delft, NL), and Thad Starner (Georgia Institute of Technology – Atlanta, US)

License © Creative Commons BY 4.0 International license
© Thomas Kosch, Kai Kunze, Christina Schneegass, and Thad Starner

The third day focused on the technical, societal, and legal challenges of cognitive sensing and interaction technologies within a longer single session. The participants discussed these challenges and ideated policies to foster benevolent use of cognitive interaction techniques. In this context, participants brainstormed which stakeholders should implement the policies. Lastly, the participants wrote an abstract about their research background, which contributed to the Dagstuhl Seminar.

7 Acknowledgments

This work was partially supported by JST Presto Grant Number JPMJPR2132. Furthermore, this work was supported by the German Research Foundation (DFG), CRC 1404: “FONDA: Foundations of Workflows for Large-Scale Scientific Data Analysis” (Project-ID 414984028).

Participants

- Esther Bosch
DLR – Braunschweig, DE
- Anna Cox
University College London, GB
- Passant Elagroudy
DFKI –
Kaiserslautern, DE
- Mustafa Hamada
Mendi Innovations –
Stockholm, SE
- Masahiko Inami
University of Tokyo, JP
- Shoya Ishimaru
Osaka Metropolitan
University, JP
- Michael Knierim
KIT – Karlsruher Institut für
Technologie, DE
- Hideki Koike
Institute of Science Tokyo, JP
- Thomas Kosch
HU Berlin, DE
- Kai Kunze
Keio University – Yokohama, JP
- Diako Mardanbegi
American University of
Beirut, LB
- Suranga Nanayakkara
National University of
Singapore, SG
- Jasmin Niess
University of Oslo, NO
- Evangelos Niforatos
TU Delft, NL
- Prasanth Sasikumar
National University of
Singapore, SG
- Christina Schneegass
TU Delft, NL
- Tanja Schultz
Universität Bremen, DE
- Thad Starner
Georgia Institute of Technology –
Atlanta, US
- Benjamin Tag
UNSW – Sydney, AU
- Steeven Villa
LMU München, DE
- Robin Welsch
Aalto University, FI
- Max L. Wilson
University of Nottingham, GB
- Anusha Withana
The University of Sydney, AU



Quantum Cryptanalysis

Gorjan Alagic^{*1}, Simona Etinski^{*2}, Stacey Jeffery^{*3}, and
Rainer Steinwandt^{*4}

1 University of Maryland – College Park, US. galagic@gmail.com

2 CWI – Amsterdam, NL. etinski.simona@gmail.com

3 CWI – Amsterdam, NL. smjeffery@gmail.com

4 University of Alabama in Huntsville, US. rs0141@uah.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25431 “Quantum Cryptanalysis”. The seminar took place as an in-person event in October 2025 and was the eighth installment of the Dagstuhl Seminar series on Quantum Cryptanalysis. This report describes the motivation and technical scope of the seminar as well as the (updated) organizational structure of this week-long event. We also include abstracts of the seminar presentations given by participants and a description of the activities of the working groups.

Seminar October 19–24, 2025 – <https://www.dagstuhl.de/25431>

2012 ACM Subject Classification Security and privacy → Cryptanalysis and other attacks

Keywords and phrases computational algebra, cryptanalysis, post-quantum cryptography, quantum algorithms, quantum resource estimation

Digital Object Identifier 10.4230/DagRep.15.10.105

1 Executive Summary

Gorjan Alagic (University of Maryland – College Park, US)

Simona Etinski (CWI – Amsterdam, NL)

Stacey Jeffery (CWI – Amsterdam, NL)

Rainer Steinwandt (University of Alabama in Huntsville, US)

License  Creative Commons BY 4.0 International license

© Gorjan Alagic, Simona Etinski, Stacey Jeffery, and Rainer Steinwandt

Motivation and technical scope

The past few years have been marked by rapid advances in quantum technologies, both from algorithmic and implementation perspectives, accompanied by a significant increase in resources devoted to the realization of fully capable quantum computers. In response to these developments, the cryptographic community has focused on designing cryptographic solutions that will remain secure and practical in the post-quantum era, once the full power of quantum computing is unleashed. A major milestone in this direction was the NIST post-quantum cryptography standardization effort, initiated in 2016 and concluded last year, which standardized several cryptographic schemes. Since the first edition – Dagstuhl Seminar “Quantum Cryptanalysis” (11381) in 2011 – in this series, we have closely followed this line of development by designing and analyzing mathematical tools to attack problems believed to be resistant to quantum attacks, intending to enable secure computation and communication in the era of large-scale quantum computers.

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Quantum Cryptanalysis, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 105–118

Editors: Gorjan Alagic, Simona Etinski, Stacey Jeffery, and Rainer Steinwandt



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In this edition, the Dagstuhl Seminar “Quantum Cryptanalysis” (25431), we placed particular emphasis on several algorithms that have seen notable improvements in recent years and that offer promising directions for further progress. Among these, we analyzed Regev’s algorithm, as well as other closely related approaches. Going a step further, we also considered more exotic algorithms based on non-abelian group actions, which provide an additional promising research direction. Furthermore, we explored a more restrictive yet potentially impactful line of research within symmetric cryptography, focusing on approaches that rely on convolution-based attacks. Finally, to maintain a strong connection to practical feasibility, we examined implementation aspects by analyzing the computational and physical resources required to realize these algorithms.

One of the central goals of this seminar series was to bring together two research communities: the quantum computing community, primarily focused on the design of quantum algorithms, and the cryptographic community, which concentrates on the construction of cryptographic primitives and the analysis of their security. The aim was to bridge gaps in understanding quantum cryptanalysis from both perspectives and to establish a shared theoretical foundation that enables meaningful communication across communities. This objective was achieved through a series of shorter, ad hoc, and on-demand talks, which served as explanatory touchpoints for topics of particular interest to the participants.

Organization

To take full advantage of the unique opportunities offered by Schloss Dagstuhl, and in line with previous editions of the seminar, ample time was reserved for discussion and collaboration. A typical day, therefore, included only two to three presentations. Before the event, the organizers contacted participants to solicit potential topics and began forming working groups. Consequently, the first day of the seminar was largely devoted to establishing these working groups and defining their technical focus.

The working groups met regularly throughout the week to discuss their respective research topics and periodically reported their progress to the entire seminar. The participant-selected working group topics were:

- Quantum algorithms for codes and lattices based on Regev’s reduction,
- Quantum cryptanalysis of non-abelian group actions,
- Algorithms for syzygies and applications,
- Quantum algorithms for factoring and computing discrete logarithms, with a focus on Regev’s algorithm,
- Topics in convolution-based quantum symmetric cryptanalysis.

Following the Dagstuhl tradition and consistent with previous seminars in the “Quantum Cryptanalysis” series, no technical program was scheduled for Wednesday afternoon. This provided participants with an opportunity to explore the surrounding area or to devote additional time to collaborative research.

With 40 participants, Schloss Dagstuhl hosted a diverse group of leading experts from around the world. A substantial portion of the attendees were graduate students. These early-career researchers benefited from close interaction with established experts, contributing to cutting-edge research discussions while gaining valuable insights to support their professional development.

Results and next steps

The working groups were once again well received, with several participants praising this seminar structure. The groups were able to make meaningful technical progress during the week, and several collaborations continued beyond the seminar. In addition, participants particularly appreciated the on-demand talks, which provided timely explanations of specific topics and helped align background knowledge across the group.

The technical presentations demonstrated that significant progress is being made in the field more broadly, underscoring the importance of research at the intersection of quantum computing and cryptanalysis. The Dagstuhl Seminar series on Quantum Cryptanalysis plays an important role in this area, and we expect this to continue as the community advances both the development of quantum computing technologies and the standardization and real-world deployment of post-quantum cryptography.

2 Table of Contents

Executive Summary

Gorjan Alagic, Simona Etinski, Stacey Jeffery, and Rainer Steinwandt 105

Overview of Talks

Convolution-based quantum cryptanalysis <i>Xavier Bonnetain, Kaveh Bashiri, and André Schrottenloher</i>	109
Quantum algorithms for codes and lattices based on Regev’s reduction <i>André Chailloux</i>	109
An Improved Quantum Algorithm for 3-Tuple Lattice Sieving <i>Lynn Engelberts</i>	110
Kuperberg’s algorithm: Opening the quantum box <i>Sean Hallgren</i>	110
QCRAM Discussion <i>Samuel E. Jaques and Christian Majenz</i>	111
Super-Quadratic Quantum Speed-Ups and Guessing Many Likely Keys <i>Julian Nowakowski</i>	111
The syzygy distinguisher <i>Hugues Randriambololona</i>	111
A Tight Quantum Algorithm for Multiple Collision Search <i>André Schrottenloher, Xavier Bonnetain, Johanna Loyer, and Yixin Shen</i>	112
Classical & Quantum Algorithms for Solving the Lattice Isomorphism Problem. <i>Wessel van Woerden</i>	112

Working groups

Quantum Cryptanalysis of Non-Abelian Group Actions <i>Giacomo Borin, Jean-François Biasse, Thomas Espitau, Alice Pellet-Mary, Christophe Petit, Simona Samardjiska, Daniel C. Smith-Tone, and Alexandre Wallet</i>	113
Quantum algorithms inspired by Regev’s reduction <i>André Chailloux, Thomas Debris-Alazard, Lynn Engelberts, Sean Hallgren, Samuel E. Jaques, Stacey Jeffery, Johanna Loyer, Michele Mosca, Julian Nowakowski, Alice Pellet-Mary, Yixin Shen, and Jean-Pierre Tillich</i>	114
Quantum algorithms for factoring and computing discrete logarithms, with a focus on Regev’s algorithm <i>Martin Ekerå, Joel Gärtner, and Gregory Kahanamoku-Meyer</i>	115
Algorithms for syzygies and applications <i>Hugues Randriambololona, Christophe Petit, Simona Samardjiska, and Jean-Pierre Tillich</i>	116
Topics in convolution-based quantum symmetric cryptanalysis <i>André Schrottenloher, Gorjan Alagic, Kaveh Bashiri, Xavier Bonnetain, Simona Etinski, Christian Majenz, and Hugues Randriambololona</i>	117

Participants	118
-------------------------------	------------

3 Overview of Talks

3.1 Convolution-based quantum cryptanalysis

Xavier Bonnetain (LORIA & INRIA – Villers-lès-Nancy, FR), Kaveh Bashiri (BSI – Bonn, DE), and André Schrottenloher (INRIA – Rennes, FR)

License © Creative Commons BY 4.0 International license

© Xavier Bonnetain, Kaveh Bashiri, and André Schrottenloher

Joint work of Kaveh Bashiri, Xavier Bonnetain, Akinori Hosoyamada, Nathalie Lang, André Schrottenloher

Main reference Kaveh Bashiri, Xavier Bonnetain, Akinori Hosoyamada, Nathalie Lang, André Schrottenloher: “Improved Quantum Linear Attacks and Application to CAST”, IACR Trans. Symmetric Cryptol., Vol. 2025(2), pp. 124–165, 2025.

URL <https://doi.org/10.46586/TOSC.V2025.I2.124-165>

Quantum algorithms are known to solve some cryptographic problems with significant advantage over classical algorithms. In this talk, we focus on the following problem: given two complex-valued Boolean functions, find the highest value of their discrete convolution. By leveraging the Quantum Fourier Transform, it is indeed possible to compute convolutions “quantumly” (with some restrictions), leading to some non-trivial quantum speedups.

After introducing the general algorithm, we look at two applications in the cryptanalysis of block ciphers, where the key-recovery can be rephrased as such a convolution problem. The first application is linear cryptanalysis, where convolutions have long been used to speedup classical key-recovery attacks, and can now be used in quantum cryptanalysis as well. The second application is differential cryptanalysis, which is less immediate and more technical. We discuss the challenges and possible further applications of this technique.

References

- 1 André Schrottenloher. *Quantum Linear Key-Recovery Attacks Using the QFT*. In *Advances in Cryptology – CRYPTO 2023*, Part V, pages 258–291, 2023.
- 2 Kaveh Bashiri, Xavier Bonnetain, Akinori Hosoyamada, Nathalie Lang, and André Schrottenloher. *Improved Quantum Linear Attacks and Application to CAST*. *IACR Transactions on Symmetric Cryptology*, 2025(2):124–165, 2025.

3.2 Quantum algorithms for codes and lattices based on Regev’s reduction

André Chailloux (INRIA – Paris, FR)

License © Creative Commons BY 4.0 International license

© André Chailloux

Joint work of André Chailloux, Jean-Pierre Tillich

Main reference André Chailloux, Jean-Pierre Tillich: “Quantum Advantage from Soft Decoders”, in Proc. of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23–27, 2025, pp. 738–749, ACM, 2025.

URL <https://doi.org/10.1145/3717823.3718319>

Shor’s quantum algorithm for integer factorization compels us to rethink currently deployed cryptographic schemes in order to make them post-quantum, that is, resistant even to adversaries equipped with a functional quantum computer. These new schemes rely on novel computational assumptions, notably those based on Euclidean lattices and error-correcting codes. In recent years, a new family of quantum algorithms targeting these lattice and code-based problems has been discovered. These algorithms are inspired by Regev’s reduction, which constitutes one of the theoretical foundations of post-quantum cryptography. This presentation will offer an overview of this family of algorithms, along with a discussion of their implications for post-quantum cryptography and, more broadly, for quantum algorithms.

3.3 An Improved Quantum Algorithm for 3-Tuple Lattice Sieving

Lynn Engelberts (CWI – Amsterdam, NL)

License  Creative Commons BY 4.0 International license
© Lynn Engelberts

Joint work of Lynn Engelberts, Yanlin Chen, Amin Shiraz Gilani, Maya-Iggy van Hoof, Stacey Jeffery, Ronald de Wolf

Main reference Lynn Engelberts, Yanlin Chen, Amin Shiraz Gilani, Maya-Iggy van Hoof, Stacey Jeffery, Ronald de Wolf: “An Improved Quantum Algorithm for 3-Tuple Lattice Sieving”, CoRR, Vol. abs/2510.08473, 2025.

URL <https://doi.org/10.48550/ARXIV.2510.08473>

The assumed hardness of the Shortest Vector Problem in high-dimensional lattices is one of the cornerstones of post-quantum cryptography. The fastest known heuristic attacks on SVP are via so-called sieving methods. While these still take exponential time in the dimension d , they are significantly faster than non-heuristic approaches, and their heuristic assumptions are verified by extensive experiments. k -Tuple sieving is an iterative method where each iteration takes as input a large number of lattice vectors of a certain norm, and produces an equal number of lattice vectors of slightly smaller norm, by taking sums and differences of k of the input vectors. Iterating these “sieving steps” sufficiently many times produces a short lattice vector. The fastest attacks (both classical and quantum) are for $k = 2$, but taking larger k reduces the amount of memory required for the attack. In this paper, we improve the quantum time complexity of 3-tuple sieving from $2^{0.3098d}$ to $2^{0.2846d}$, using a two-level amplitude amplification aided by a preprocessing step that associates the given lattice vectors with nearby “center points” to focus the search on the neighborhoods of these center points. Our algorithm uses $2^{0.1887d}$ classical bits and QCRAM bits, and $2^{o(d)}$ qubits. This is the fastest known quantum algorithm for SVP when total memory is limited to $2^{0.1887d}$.

3.4 Kuperberg’s algorithm: Opening the quantum box

Sean Hallgren (Pennsylvania State University – University Park, US)

License  Creative Commons BY 4.0 International license
© Sean Hallgren

Main reference Greg Kuperberg: “A Subexponential-Time Quantum Algorithm for the Dihedral Hidden Subgroup Problem”, SIAM J. Comput., Vol. 35(1), pp. 170–188, 2005.

URL <https://doi.org/10.1137/S0097539703436345>

Kuperberg’s algorithm is a subexponential time algorithm for the dihedral hidden subgroup problem. This tutorial presented the algorithm in detail, also serving as an example of what a quantum algorithm looks like, for those new to the field.

(Talk abstract written by S. Jeffery.)

3.5 QCRAM Discussion

Samuel E. Jaques (University of Waterloo, CA) and Christian Majenz (Technical University of Denmark – Lyngby, DK)

License © Creative Commons BY 4.0 International license
© Samuel E. Jaques and Christian Majenz

In this discussion, we first clarified the distinction between different types of quantum and classical memory, and then had a lively discussion about which models are physically realistic. We reached a general consensus that it is important to always clarify the specific memory resources your quantum algorithms require.

(Talk abstract written by S. Jeffery.)

3.6 Super-Quadratic Quantum Speed-Ups and Guessing Many Likely Keys

Julian Nowakowski (Ruhr-Universität Bochum, DE)

License © Creative Commons BY 4.0 International license
© Julian Nowakowski
Joint work of Timo Glaser, Alexander May, Julian Nowakowski
Main reference Timo Glaser, Alexander May, Julian Nowakowski: “Super-Quadratic Quantum Speed-Ups and Guessing Many Likely Keys”, 2023.
URL <https://eprint.iacr.org/2023/797>

In this talk, we study the fundamental problem of guessing cryptographic keys, drawn from some probability distribution D . The optimal classical algorithm enumerates keys in decreasing order of likelihood. The optimal quantum algorithm, due to Montanaro (2011), is a sophisticated Grover search.

We give the first tight analysis for Montanaro’s algorithm, showing that its runtime is $2^{H_{2/3}(D)/2}$, where H_α denotes Renyi entropy with parameter α . Interestingly, this is a direct consequence of an information theoretic result called Arian’s Inequality (1995) – which has so far been missed in the cryptographic community – that tightly bounds the runtime of classical key guessing by $2^{H_{1/2}(D)}$.

Since $H_{2/3}(D) < H_{1/2}(D)$ for every non-uniform distribution D , we thus obtain a super-quadratic quantum speed-up over classical key guessing.

Additionally, we provide the first thorough analysis of guessing in a multi-key setting.

3.7 The syzygy distinguisher

Hugues Randriambololona (ANSSI – Paris, FR)

License © Creative Commons BY 4.0 International license
© Hugues Randriambololona
Main reference Hugues Randriambololona: “The syzygy Distinguisher”, in Proc. of the Advances in Cryptology – EUROCRYPT 2025 – 44th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Madrid, Spain, May 4-8, 2025, Proceedings, Part VI, Lecture Notes in Computer Science, Vol. 15606, pp. 324–354, Springer, 2025.
URL https://doi.org/10.1007/978-3-031-91095-1_12

In this talk I introduced higher modules of syzygies for codes, and associated notions (minimal resolution, graded Betti numbers). I then explained how, applied to a suitable shortening of the dual code, these invariants allow to distinguish binary Goppa codes (as they appear

in the McEliece cryptosystem) from random codes. The asymptotic complexity of this distinguisher is (slightly) subexponential in the error-correcting capability, which is better than any previous cryptanalytic attempt against the McEliece cryptosystem (either targetting the key or the messages). Last, I briefly discussed work in progress and recent advances that occurred after this work was presented at Eurocrypt 2025.

3.8 A Tight Quantum Algorithm for Multiple Collision Search

André Schrottenloher (INRIA – Rennes, FR), Xavier Bonnetain (LORIA & INRIA – Villers-lès-Nancy, FR), Johanna Loyer (INRIA Saclay – Palaiseau, FR), and Yixin Shen (INRIA – Rennes, FR)

License © Creative Commons BY 4.0 International license

© André Schrottenloher, Xavier Bonnetain, Johanna Loyer, and Yixin Shen

Main reference Xavier Bonnetain, Johanna Loyer, André Schrottenloher, Yixin Shen: “A Tight Quantum Algorithm for Multiple Collision Search”, 2025.

URL <https://eprint.iacr.org/2025/1690>

Searching for collisions in random functions is a fundamental computational problem, with many applications in symmetric and asymmetric cryptanalysis. When one searches for a *single* collision, the known quantum algorithms match the query lower bound. This is not the case for the problem of finding *multiple* collisions, despite its regular appearance as a sub-component in sieving-type algorithms.

At EUROCRYPT 2019, Liu and Zhandry gave a query lower bound $\Omega(2^{m/3+2k/3})$ for finding 2^k collisions in a random function with m -bit output. At EUROCRYPT 2023, Bonnetain et al. gave a quantum algorithm matching this bound for a large range of m and k , but not all admissible values. Like many previous collision-finding algorithms, theirs is based on the MNRS quantum walk framework, but it *chains* the walks by reusing the state after outputting a collision.

In this talk, we give a new algorithm that tackles the remaining non-optimal range, closing the problem. Our algorithm is tight (up to a polynomial factor) in queries, and also in time under a quantum RAM assumption. The idea is to extend the chained walk to a regime in which several collisions are returned at each step, and the “walks” themselves only perform a single diffusion layer.

3.9 Classical & Quantum Algorithms for Solving the Lattice Isomorphism Problem.

Wessel van Woerden (PQShield – Amsterdam, NL)

License © Creative Commons BY 4.0 International license

© Wessel van Woerden

In this talk I will survey the currently known algorithms for solving the Lattice Isomorphism Problem (LIP) and its module version (Module-LIP). The latter case also involves some quantum steps, which can optionally be removed with additional assumptions and heuristics.

4 Working groups

4.1 Quantum Cryptanalysis of Non-Abelian Group Actions

Giacomo Borin (IBM Research Europe – Zürich, CH), Jean-François Biasse (University of South Florida – Tampa, US), Thomas Espitau (PQShield – Paris, FR), Alice Pellet-Mary (University of Bordeaux, FR), Christophe Petit (UL – Brussels, BE), Simona Samardjiska (Radboud University Nijmegen, NL), Daniel C. Smith-Tone (NIST – Gaithersburg, US), and Alexandre Wallet (PQShield – Paris, FR)

License © Creative Commons BY 4.0 International license

© Giacomo Borin, Jean-François Biasse, Thomas Espitau, Alice Pellet-Mary, Christophe Petit, Simona Samardjiska, Daniel C. Smith-Tone, and Alexandre Wallet

The working group tried to explore the potential of applying quantum algorithms to cryptanalysis of the discrete logarithm problem on non-abelian group actions (also called vectorization problem). Two options were considered:

1. Applying the Kuperberg algorithm to the linear code equivalence problem. This approach did not yield any significant results.
2. Combining the reduction from [2] and the quantum algorithm from [1]. This approach was more promising, with some partial results obtained.

The main ideas were to rewrite the Principal Ideal Problem as a Borel hidden subgroup problem efficiently solvable with a generalization of the algorithm from [1]. Even if initial results were encouraging, the structure of the secret did not seem to be compatible with the information retrieved by the quantum algorithm.

Further work is needed to determine if this approach can be made to work.

References

- 1 Chen, Mingjie, Muhammad Imran, Gábor Ivanyos, Péter Kutas, Antonin Leroux, and Christophe Petit. “Hidden stabilizers, the isogeny to endomorphism ring problem and the cryptanalysis of pSIDH.” In *International Conference on the Theory and Application of Cryptology and Information Security*, pp. 99-130. Singapore: Springer Nature Singapore, 2023.
- 2 Cheignard, Clémence, Guilhem Mureau, Thomas Espitau, Alice Pellet-Mary, Heorhii Pliatsok, and Alexandre Wallet. “A reduction from Hawk to the principal ideal problem in a quaternion algebra.” In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 154-183. Cham: Springer Nature Switzerland, 2025.

4.2 Quantum algorithms inspired by Regev’s reduction

André Chailloux (INRIA – Paris, FR), Thomas Debris-Alazard (Ecole Polytechnique – Palaiseau, FR & Inria Saclay – Palaiseau, FR), Lynn Engelberts (CWI – Amsterdam, NL), Sean Hallgren (Pennsylvania State University – University Park, US), Samuel E. Jaques (University of Waterloo, CA), Stacey Jeffery (CWI – Amsterdam, NL), Johanna Loyer (INRIA Saclay – Palaiseau, FR), Michele Mosca (University of Waterloo, CA), Julian Nowakowski (Ruhr-Universität Bochum, DE), Alice Pellet-Mary (University of Bordeaux, FR), Yixin Shen (INRIA – Rennes, FR), and Jean-Pierre Tillich (INRIA – Paris, FR)

License © Creative Commons BY 4.0 International license

© André Chailloux, Thomas Debris-Alazard, Lynn Engelberts, Sean Hallgren, Samuel E. Jaques, Stacey Jeffery, Johanna Loyer, Michele Mosca, Julian Nowakowski, Alice Pellet-Mary, Yixin Shen, and Jean-Pierre Tillich

There is a recent family of quantum algorithm based on Regev’s reduction that has gained a lot of attention recently. In its modern formulation, this is a reduction between the Inhomogeneous Short Integer Solution Problem (ISIS) to the $|S\text{-LWE}\rangle$ which is a variant of LWE where the error is in quantum superposition. For constructing quantum algorithms, this framework has first been presented by Chen, Liu and Zhandry and there have since been many works trying to leverage this reduction for solving code and lattice-based problems which are classically intractable. This approach has been quite successful and we have several examples of powerful quantum algorithms. However, the problems that can be attacked are still far from real problems used in post-quantum cryptography.

The idea of this working group was to have a general discussion on this family of algorithms. The main question was to determine whether these can have a strong impact on quantum cryptanalysis and therefore on the foundations of current post-quantum cryptography. We touched on the following questions:

- Find structured codes/lattices for which these algorithms yield a quantum advantage.
- Apart from Gaussian Elimination and the Arora-Ge algorithm, do we have access to other “generic” algorithms for LWE that we can use?
- Can we “quantize” these algorithms so that they are more efficient in the presence of quantum noise?
- Do we have access to more tools from quantum information theory than just (partial) unambiguous measurements?
- Can we say something about the classical hardness that these problems have or can we dequantize these quantum algorithms?
- How far are we from attacking a system like DILITHIUM with these algorithms?

The conclusion of this working group is that while the approach is promising, we don’t yet have examples where this can be directly used for lattice-based problems and that one requires a more in-depth understanding of the $|S\text{-LWE}\rangle$ and the differences with the LWE problem in order to fully take advantage of this approach.

4.3 Quantum algorithms for factoring and computing discrete logarithms, with a focus on Regev’s algorithm

Martin Ekerå (KTH Royal Institute of Technology – Stockholm, SE & Swedish NCSA, SE), Joel Gärtner (KTH Royal Institute of Technology – Stockholm, SE), and Gregory Kahanamoku-Meyer (MIT – Cambridge, US)

License © Creative Commons BY 4.0 International license
© Martin Ekerå, Joel Gärtner, and Gregory Kahanamoku-Meyer

The discussion in the breakout group first centered on the practicability of Regev’s algorithm [10, 3]. The conclusion of the group was that although the new parallel spooky pebbling-based approach [7] reduces the circuit size and depth compared to earlier Fibonacci-based approaches [9, 8], it is still hard to see a niche for Regev’s algorithm where it would be of practical interest unless further optimizations are found or non-computational space is cheap.

The group considered various metrics that may be useful in comparing quantum algorithms when coarse-grained metrics such as that used in [4] no longer suffice to distinguish the best implementation, including but not limited to the space usage, depth, gate count and spacetime volume, both per-run and overall, and both logically and physically. The group agreed that in the short to medium term, algorithms and implementation techniques that achieve a small space footprint, most notably [11, 5, 2, 1, 6], are currently arguably the best option. In the long term, what matters most is arguably the spacetime volume across all runs (qubits times depth), and here too it appears hard for Regev’s algorithm to compete given the number of runs it requires (not least when considering the overhead of quantum error correction).

Furthermore, the group discussed options for parallelizing the pre-Regev algorithms when a similar amount of space is available as that required by Regev’s algorithm. With respect to parallelization, the group agreed that limited parallelization in [6] may be beneficial, as opposed to sequentially processing the many small primes. This would result in fairly substantial depth reduction at the expense of only a small increase in the space usage.

Finally, the group considered options for further reducing the space required to store the control registers in [1, 6], for instance via recycling some of the qubits. However, the conclusion of the group was that it would take new ideas to achieve such a space reduction without incurring a space blowup in other parts of the circuit.

References

- 1 C. Chevalignard, P.-A. Fouque, and A. Schrottenloher. *Reducing the Number of Qubits in Quantum Factoring*. In: Advances in Cryptology – CRYPTO 2025 (LNCS 16001), Springer, pages 384–415, 2025. doi:10.1007/978-3-032-01878-6_13
- 2 M. Ekerå. *On post-processing in the quantum algorithm for computing short discrete logarithms*. Des. Codes Cryptogr., 88(11):2313–2335, Springer, 2020. doi:10.1007/s10623-020-00783-2
- 3 M. Ekerå and J. Gärtner. *Extending Regev’s Factoring Algorithm to Compute Discrete Logarithms*. In: Post-Quantum Cryptography — PQCrypto 2024 (LNCS 14772), Springer, 2024, pp. 211–242. doi:10.1007/978-3-031-62746-0_10
- 4 M. Ekerå and J. Gärtner. *A High-Level Comparison of State-of-the-Art Quantum Algorithms for Breaking Asymmetric Cryptography*. IACR Comm. Cryptol., 2(1):33, 2025. doi:10.62056/ayzobjb0kr
- 5 M. Ekerå and J. Håstad. *Quantum Algorithms for Computing Short Discrete Logarithms and Factoring RSA Integers*. In: Post-Quantum Cryptography — PQCrypto 2017 (LNCS 10346), Springer, 2017, pp. 347–363. doi:10.1007/978-3-319-59879-6_20

- 6 C. Gidney. *How to factor 2048 bit RSA integers with less than a million noisy qubits*. ArXiv preprint arXiv:2505.15917, 2025. doi:10.48550/arXiv.2505.15917
- 7 G. D. Kahanamoku-Meyer, S. Ragavan, and K. Van Kirk. *Parallel Spooky Pebbling Makes Regev Factoring More Practical*. ArXiv preprint arXiv:2510.08432, 2025. doi:10.48550/arXiv.2510.08432
- 8 S. Ragavan. *Regev Factoring Beyond Fibonacci: Optimizing Prefactors*. Cryptology ePrint Archive, Paper 2024/636, 2024. url:https://ia.cr/2024/636
- 9 S. Ragavan and V. Vaikuntanathan. *Space-Efficient and Noise-Robust Quantum Factoring*. In: *Advances in Cryptology — CRYPTO 2024 (LNCS 14925)*, Springer, 2024, pp. 107–140. doi:10.1007/978-3-031-68391-6_4
- 10 O. Regev. *An Efficient Quantum Factoring Algorithm*. J. ACM, 72(1):10, ACM, 2025. doi:10.1145/3708471
- 11 P. W. Shor. *Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer*. SIAM J. Comput., 26(5):1484–1509, 1997. doi:10.1137/S0097539795293172

4.4 Algorithms for syzygies and applications

Hugues Randriambololona (ANSSI – Paris, FR), Christophe Petit (UL – Brussels, BE), Simona Samardjiska (Radboud University Nijmegen, NL), and Jean-Pierre Tillich (INRIA – Paris, FR)

License © Creative Commons BY 4.0 International license

© Hugues Randriambololona, Christophe Petit, Simona Samardjiska, and Jean-Pierre Tillich

A first goal of this working group was to help participants get more familiarity or intuition with higher modules of syzygies. We then presented various algorithms to compute syzygies:

- the iterative approach, following the definitions
- a direct approach through Koszul cohomology.

We then discussed possible techniques (classical or quantum) that could speed-up the underlying linear algebra:

- block Wiedemann type techniques
- the HHL algorithm.

Last we tried to identify other possible applications of syzygies in cryptanalysis.

4.5 Topics in convolution-based quantum symmetric cryptanalysis

André Schrottenloher (INRIA – Rennes, FR), Gorjan Alagic (University of Maryland – College Park, US), Kaveh Bashiri (BSI – Bonn, DE), Xavier Bonnetain (LORIA & INRIA – Villers-lès-Nancy, FR), Simona Etinski (CWI – Amsterdam, NL), Christian Majenz (Technical University of Denmark – Lyngby, DK), and Hugues Randriambololona (ANSSI – Paris, FR)

License © Creative Commons BY 4.0 International license

© André Schrottenloher, Gorjan Alagic, Kaveh Bashiri, Xavier Bonnetain, Simona Etinski, Christian Majenz, and Hugues Randriambololona

Main reference Kaveh Bashiri, Xavier Bonnetain, Akinori Hosoyamada, Nathalie Lang, André Schrottenloher: “Improved Quantum Linear Attacks and Application to CAST”, 2024.

URL <https://eprint.iacr.org/2024/1992>

This working group explored open questions related to block cipher cryptanalysis using the quantum convolution algorithm (QCA). This is a family of statistical attacks where the statistic of interest is expressed as a convolution, and used to distinguish right and wrong key guesses. We targeted in particular two problems.

The first problem was the use of convolution-based cryptanalysis in the case of zero-correlation linear attacks. In these linear attacks, the statistic for the right key guess becomes zero. Unfortunately in the quantum setting, it is the amplitude over the key itself which becomes zero, which is useless. We tried different manipulations of the QCA output which would have improved this situation, especially using independent zero-correlation statistics. Unfortunately our attempts in this direction failed.

The second problem was to make use of precomputations in the attack. Indeed, the computation of the QCA, at least in linear cryptanalysis, is not the dominating cost of the attack. We tried different approaches to gain a better advantage by pre-amplifying the output of the QCA. One of these attempts worked: we obtained a block cipher structure on which a non-trivial attack could be mounted. This attack can be seen roughly as a combination of Meet-in-the-middle cryptanalysis with linear cryptanalysis, which warrants further investigation.

Participants

- Gorjan Alagic
University of Maryland –
College Park, US
- Daniel C. Apon
Anduril Industries –
Costa Mesa, US
- Kaveh Bashiri
BSI – Bonn, DE
- Jean-François Biasse
University of South Florida –
Tampa, US
- Xavier Bonnetain
LORIA & INRIA –
Villers-lès-Nancy, FR
- Giacomo Borin
IBM Research Europe –
Zürich, CH
- Brennon Brimhall
Anduril Industries –
Costa Mesa, US
- André Chailloux
INRIA – Paris, FR
- Yanlin Chen
University of Maryland –
College Park, US
- Thomas Debris-Alazard
Ecole Polytechnique – Palaiseau,
FR & Inria Saclay –
Palaiseau, FR
- Martin Ekerå
KTH Royal Institute of
Technology – Stockholm, SE &
Swedish NCSA, SE
- Lynn Engelberts
CWI – Amsterdam, NL
- Thomas Espitau
PQShield – Paris, FR
- Simona Etinski
CWI – Amsterdam, NL
- Joel Gärtner
KTH Royal Institute of
Technology – Stockholm, SE
- Amin Shiraz Gilani
University of Maryland –
College Park, US
- Sean Hallgren
Pennsylvania State University –
University Park, US
- Kelsey Jackson
University of Maryland –
College Park, US
- Samuel E. Jaques
University of Waterloo, CA
- Stacey Jeffery
CWI – Amsterdam, NL
- Gregory Kahanamoku-Meyer
MIT – Cambridge, US
- Susanna Kirchhoff
Forschungszentrum Jülich, DE
- Peter Kristel
Cyberagentur – Halle, DE
- Johanna Loyer
INRIA Saclay – Palaiseau, FR
- Laura Maddison
University of Calgary, CA
- Christian Majenz
Technical University of Denmark
– Lyngby, DK
- Michele Mosca
University of Waterloo, CA
- Julian Nowakowski
Ruhr-Universität Bochum, DE
- Alice Pellet-Mary
University of Bordeaux, FR
- Christophe Petit
UL – Brussels, BE
- Hugues Randriambololona
ANSSI – Paris, FR
- Simona Samardjiska
Radboud University
Nijmegen, NL
- André Schrottenloher
INRIA – Rennes, FR
- Yixin Shen
INRIA – Rennes, FR
- Manasi Shingane
University of Maryland –
College Park, US
- Daniel C. Smith-Tone
NIST – Gaithersburg, US
- Jean-Pierre Tillich
INRIA – Paris, FR
- Wessel van Woerden
PQShield – Amsterdam, NL
- Alexandre Wallet
PQShield – Paris, FR
- Bo-Yin Yang
Academia Sinica – Taipei, TW



Deep Continual Learning in the Foundation Model Era

Christopher Kanan^{*1}, Martin Mundt^{*2}, Tinne Tuytelaars^{*3},
Joost van de Weijer^{*4}, and Timm Felix Hess^{†5}

1 University of Rochester, US. chriskanan@gmail.com

2 Universität Bremen, DE. mundtm@uni-bremen.de

3 KU Leuven, BE. tinne.tuytelaars@esat.kuleuven.be

4 Computer Vision Center – Barcelona, ES. joost@cvc.uab.es

5 KU Leuven, BE. TimmFelix.Hess@esat.kuleuven.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25432 “Deep Continual Learning in the Foundation Model Era”. This seminar brought together 23 researchers to discuss research at the intersection of continual learning and foundation models. The discussion centered on the major challenges arising from continual training of foundation models, including the need for new benchmarks, new opportunities for memory-based continual learning, emerging application domains, and the development of efficient metrics to quantify forgetting of foundation model knowledge. In addition, the report contains a summary of the talks of the participants.

Seminar October 19–24, 2025 – <https://www.dagstuhl.de/25432>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Neural networks; Computing methodologies → Learning settings; Computing methodologies → Learning paradigms; Computing methodologies → Bio-inspired approaches; Computing methodologies → Learning latent representations; Computing methodologies → Computer vision; Computing methodologies → Natural language processing; Computing methodologies → Distributed artificial intelligence

Keywords and phrases continual learning, deep learning, foundation models

Digital Object Identifier 10.4230/DagRep.15.10.119

1 Executive Summary

Christopher Kanan (University of Rochester, US, chriskanan@gmail.com)

Martin Mundt (Universität Bremen, DE, mundtm@uni-bremen.de)

Tinne Tuytelaars (KU Leuven, BE, tinne.tuytelaars@esat.kuleuven.be)

Joost van de Weijer (Computer Vision Center – Barcelona, ES, joost@cvc.uab.es)

License  Creative Commons BY 4.0 International license

© Christopher Kanan, Martin Mundt, Tinne Tuytelaars, and Joost van de Weijer

This summary presents the main outcomes of the Dagstuhl Seminar “Deep Continual Learning in the Foundation Model Era” (25432). Foundation models have transformed artificial intelligence and generated major socio-economic impact. However, they face significant challenges: the prevailing training paradigm, which relies on retraining models from scratch as new data arrives, is unsustainable, models must also be adapted to better align with human values and mitigate biases, and adaptation often leads to catastrophic forgetting. The field of deep continual learning focuses on enabling systems to acquire knowledge over time

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Deep Continual Learning in the Foundation Model Era, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 119–134
Editors: Christopher Kanan, Martin Mundt, Tinne Tuytelaars, Joost van de Weijer, and Timm Felix Hess



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

from non-stationary data streams. It provides theoretical and methodological tools that can help address key limitations of foundation models, including efficient updating, adaptation at scale, alignment with human values, and bias mitigation.

The objective of the seminar was to bring together world-class researchers in the fields of deep continual learning and foundation model training to examine their interplay, identify key challenges, and outline promising directions for future research.

As a major outcome of the seminar, the following promising research directions were identified:

- New memory usage paradigms are emerging for continual learning in foundation models. Approaches that combine parametric learning with in-context learning are a promising path toward rapid adaptation and experience accumulation. In-context learning reduces the need for costly and forgetting-prone parameter updates, while periodic consolidation of in-context knowledge through distillation into model parameters may limit the computational overhead associated with growing context.
- New methods are needed for the efficient evaluation of catastrophic forgetting in foundation models during adaptation, as full-model evaluation is computationally prohibitive in iterative adaptation settings.
- Model scale, pre-training, and architectural choices fundamentally alter the stability-plasticity trade-off. Systematic studies across model scales and objectives are needed to understand when scale inherently mitigates forgetting and when explicit continual learning mechanisms remain necessary.
- Continual learning for foundation models requires new benchmarks that go beyond synthetic task sequences and better reflect real-world usage. Promising benchmark domains include personalization over time, multi-agent interaction, model self-improvement loops, and unlearning for privacy or safety. These benchmarks should impose realistic compute constraints.

2 Table of Contents

Executive Summary

Christopher Kanan, Martin Mundt, Tinne Tuytelaars, and Joost van de Weijer . . . 119

Overview of Talks

Three forms of continual learning from foundation models <i>Rahaf Aljundi</i>	123
How green is continual adaptation of Foundation Models, really? <i>Andrew D. Bagdanov</i>	123
Modular Continual Learning <i>Lucas Caccia</i>	123
The Myth of Catastrophic Forgetting <i>Antonio Carta</i>	124
Toward “Model-based” Continual Learning <i>Laurent Charlin</i>	124
Three questions/issues on continual learning <i>Jonghyun Choi</i>	124
XAI in incremental learning <i>Barbara Hammer</i>	125
Adapting to the Unknown: Novel Class Discovery, Open-Vocabulary Learning, & Beyond <i>Tyler Hayes</i>	125
Is there a benefit in joint training? <i>Timm Felix Hess</i>	125
The Death of “AGI” and the Rebirth of Continual Learning <i>Christopher Kanan</i>	126
Rethinking Efficiency in Continual Learning <i>Dhiresha Kudithipudi</i>	126
Continual Learning on Vision-Language Pretrained Models <i>Xialei Liu</i>	127
Continual, Decentralized Compositionality for Sustainable Artificial Intelligence <i>Vincenzo Lomonaco</i>	127
Modularity for Continual Robot Learning <i>Jorge Mendez-Mendez</i>	128
What does it mean to accumulate knowledge in lifelong (machine) learning? <i>Martin Mundt</i>	128
CoPeP: Continual Pretraining for Protein Language Models <i>Darshan Patil</i>	128
A Metacognition Lens on Continual Foundation Model Learning: Progress & Challenges <i>Ameya Prabhu</i>	129

Beyond Zero-Shot Generalization: Strengthening Vision-Language Models with Adaptation & Personalization	
<i>Elisa Ricci</i>	129
Some reflections on Deep Continual Learning in the foundation model era	
<i>Tinne Tuytelaars</i>	130
Do we need test-time adaptation for VLMs? The side-effects of unlearning	
<i>Bartłomiej Twardowski</i>	130
How to look at forgetting?	
<i>Gido van de Ven</i>	131
The Butterfly Effect of Minimal Updates on Generative Visual Models	
<i>Joost van de Weijer</i>	131
Continual Learning for Multi-modal Human-centric Applications	
<i>Liyuan Wang</i>	131
Working Groups	
Working Group on Main Changes for Continual Learning in the Foundation Model Era	132
Working Group on Memory in Continual Learning in the Foundation Model Era	132
Working Group on Forgetting in the Foundation Model Era	132
Working Group on Benchmarks, Setups, and Evaluation in the Foundation Model Era	133
Participants	134

3 Overview of Talks

3.1 Three forms of continual learning from foundation models

Rahaf Aljundi (Toyota Motor Europe – Zaventem, BE)

License © Creative Commons BY 4.0 International license
© Rahaf Aljundi

Continual learning enables models to adapt to streaming data while retaining knowledge, but traditional parameter updates risk catastrophic forgetting. Foundation models exhibit in-context learning capability, offering a complementary approach to adaptation. I propose a system combining a core reasoning model with short- and long-term memory. Short-term memory captures immediate context and reasoning, while long-term memory aggregates these for slow, informed updates. This architecture supports rapid adaptation and sustained improvement without full retraining. Empirical insights from presented work show in-context learning outperforming textual memory and fine-tuning, suggesting memory-centric strategies could further enhance performance, including in vision-language tasks.

3.2 How green is continual adaptation of Foundation Models, really?

Andrew D. Bagdanov (University of Florence, IT)

License © Creative Commons BY 4.0 International license
© Andrew D. Bagdanov

As research in Continual Learning continues gaining momentum, it is essential to understand its implications in the Foundation Model Era. The typically massive scale of foundation models demands efficient methods for continual adaptation, which has prompted a shift from monolithic approaches to more modular techniques such as prompt learning and adapter-based methods. While energy efficiency is often touted as an advantage of Continual Learning, a systematic comparative analysis of these methods regarding their efficiency-accuracy trade-offs is still missing in the literature. In this talk, I will explore the architectural differences between representative monolithic and modular approaches, and present an empirical evaluation of their energy consumption during training and inference. By uncovering the true energy costs of these methods, we hope to better assess their sustainability and impact on future AI research and practice.

3.3 Modular Continual Learning

Lucas Caccia (Microsoft Research – Montréal, CA)

License © Creative Commons BY 4.0 International license
© Lucas Caccia

In this talk I present different approaches to modular continual learning. Modularity in parameter, model and data space offer different tradeoffs, each with key properties that are required to enable long term, efficient, continual learning systems.

3.4 The Myth of Catastrophic Forgetting

Antonio Carta (University of Pisa, IT)

License  Creative Commons BY 4.0 International license
 Antonio Carta

In this talk I discuss some limitations of continual learning setups, and how they impacted the design space of continual learning methods. Then, I highlight some open challenges such as robustness to noise, better measures of representation forgetting, and continual optimization.

3.5 Toward “Model-based” Continual Learning

Laurent Charlin (HEC Montréal, CA)

License  Creative Commons BY 4.0 International license
 Laurent Charlin

Joint work of Laurent Charlin, Yipeng Zhang, Hafez Ghaemi, Jungyoon Lee, Shahab Bakhtiari, Eilif B. Mülle
Main reference Anonymous: “Self-Supervised Learning from Structural Invariance”, in Proc. of the Submitted to The Fourteenth International Conference on Learning Representations, 2025.
URL <https://openreview.net/forum?id=r3JUDAYjIH>

We suggest using “natural” processes as a source of continuous data. Such data has interesting properties, including temporal consistency, multi-level periodicity, and the appearance of novel elements. We also observe that “close by” data (for example, two photos of the same landscape taken at slightly different times) might be modelled as being generated from two representations that change sparsely. We instantiate this idea for learning image representation (using SSL) from natural pairs, demonstrating that this type of continuous data is beneficial for representation learning. We also propose a model inspired by the generative data process.

3.6 Three questions/issues on continual learning

Jonghyun Choi (Seoul National University, KR)

License  Creative Commons BY 4.0 International license
 Jonghyun Choi

Joint work of Jonghyun Choi, Minhyuk Seo, Tinne Tuytelaars
Main reference Minhyuk Seo, Hyunseo Koh, Jonghyun Choi: “Budgeted Online Continual Learning by Adaptive Layer Freezing and Frequency-based Sampling”, in Proc. of the The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025, OpenReview.net, 2025.
URL <https://openreview.net/forum?id=dOakHmsjRX>

I start with three questions: What is continual learning (CL) beyond class-incremental learning (CIL)? Are online CIL methods being compared fairly? And why does continual-like training improve multimodal foundation models? I introduce CIL setups we proposed so far. I then argue that considering computational cost among CIL methods is crucial for fair ground, introducing our computationally efficient CIL method. For the third question, I made a connection between federated learning (spatially distributed multitask learning) and continual learning (temporally distributed multitask learning). Finally, I introduce our recent work for federated learning with heterogeneous models for personalization.

3.7 XAI in incremental learning

Barbara Hammer (Universität Bielefeld, DE)

License © Creative Commons BY 4.0 International license
© Barbara Hammer

How can persons deal with continuously learning models? I will argue that explainability is a crucial issue to enable persons to understand how and why a model has changed. I will have a glimpse at incremental explanations and explanation of drift, and I will argue that we need to move towards XAI technologies which also address interactions when dealing with multimodal foundation models.

3.8 Adapting to the Unknown: Novel Class Discovery, Open-Vocabulary Learning, & Beyond

Tyler Hayes (Georgia Institute of Technology – Atlanta, US)

License © Creative Commons BY 4.0 International license
© Tyler Hayes

In this talk, I argue that there are two complementary paradigms for enabling adaptation in neural networks: dynamic adaptation and static adaptation. Dynamic adaptation methods adapt the model on the fly by exploiting a stream of incoming data, as in novel class discovery. In contrast, static adaptation methods prepare the model in advance for unavailable semantic concepts, often by leveraging external knowledge sources such as natural language (e.g., open-vocabulary learning). I then present our PANDAS method that uses prototypes and a distance-based classifier to adapt to novel classes, before presenting our SHiNe method that fuses class taxonomy information into a classifier to improve open-vocabulary performance. I then conclude with my vision for the future of novel concept discovery and learning.

3.9 Is there a benefit in joint training?

Timm Felix Hess (KU Leuven, BE)

License © Creative Commons BY 4.0 International license
© Timm Felix Hess

Joint work of Timm Felix Hess, Abhishek Jha, Gido M van de Ven, Tinne Tuytelaars

We analyze the challenges posed by partial observability when training on sequential data, examining its impact beyond catastrophic forgetting. Preliminary results indicate that the partial observability of sequential observations inhibits learning, causing a performance deficit relative to a jointly trained model. We position this perspective as essential to Continual Learning, where achieving joint-model performance typically is considered the ultimate goal. We also link it to the open problem of how (foundation) models can effectively integrate separate, sequential observations to build a generalized knowledge base.

3.10 The Death of “AGI” and the Rebirth of Continual Learning

Christopher Kanan (University of Rochester, US)

License  Creative Commons BY 4.0 International license
© Christopher Kanan

For years, continual learning was treated as a prerequisite for AGI. Today, many claim AGI is already here, at least under flexible, labor-centric definitions. Either way, the old continual learning agenda no longer fits. This talk proposes a reset that aligns with two concrete futures: (1) economic automation, where frontier models stay useful through fast, compute-aware updates; and (2) human-like synthetic minds, where systems form, consolidate, and reuse memories over long horizons. Minimizing forgetting is the wrong goal. The right objective is to maximize retained competence and forward transfer per unit of compute. I will share simple, high-leverage mechanisms that move the needle in practice: structured replay with wake and sleep, stability-gap mitigation for rapid adaptation, and selective rehearsal that progresses from easy to hard. These ideas improve production-style foundation models and lay the groundwork for memory-centric agents. The takeaway is clear: storage is cheap, data is abundant, compute is precious. Continual learning should be the default operating mode for models that matter.

3.11 Rethinking Efficiency in Continual Learning

Dhiresha Kudithipudi (University of Texas – San Antonio, US)

License  Creative Commons BY 4.0 International license
© Dhiresha Kudithipudi

In this talk, we present the need for thinking across the stack to enable efficient continual learning that goes beyond performance improvements. We explore how integrating insights across the neuroscience, algorithm, architecture, and system layers offers a unique opportunity to design efficient systems. By distilling shared principles from hardware and software, we can develop more effective strategies for adaptation, replay, and regularization. This cross-stack co-design reveals how to build foundation models that are robust, compute, energy, and latency efficient.

3.12 Continual Learning on Vision-Language Pretrained Models

Xialei Liu (Nankai University – Tianjin, CN)

- License** © Creative Commons BY 4.0 International license
© Xialei Liu
- Joint work of** Xusheng, Cao, Linglan Huang
- Main reference** Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, Xialei Liu: “Mind the Gap: Preserving and Compensating for the Modality Gap in CLIP-Based Continual Learning”, CoRR, Vol. abs/2507.09118, 2025.
URL <https://doi.org/10.48550/ARXIV.2507.09118>
- Main reference** Xusheng Cao, Haori Lu, Linlan Huang, Xialei Liu, Ming-Ming Cheng: “Generative Multi-modal Models are Good Class-Incremental Learners”, in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 28706–28717, IEEE, 2024.
URL <https://doi.org/10.1109/CVPR52733.2024.02712>
- Main reference** Linlan Huang, Xusheng Cao, Haori Lu, Xialei Liu: “Class-Incremental Learning with CLIP: Adaptive Representation Adjustment and Parameter Fusion”, in Proc. of the Computer Vision – ECCV 2024 – 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LIV, Lecture Notes in Computer Science, Vol. 15112, pp. 214–231, Springer, 2024.
URL https://doi.org/10.1007/978-3-031-72949-2_13
- Main reference** Xusheng Cao, Haori Lu, Linlan Huang, Fei Yang, Xialei Liu, Ming-Ming Cheng: “Knowledge Graph Enhanced Generative Multi-modal Models for Class-Incremental Learning”, CoRR, Vol. abs/2503.18403, 2025.
URL <https://doi.org/10.48550/ARXIV.2503.18403>

Continual learning is one of the key capabilities of next-generation artificial intelligence systems. It aims to enable systems to continuously acquire new knowledge in dynamically changing environments while avoiding catastrophic forgetting of previously learned knowledge, thereby mimicking the way humans learn. This talk explores continual learning in multimodal models based on vision-language pre-training frameworks, addressing current challenges from both the perspective of discriminative models (e.g., CLIP) and generative models (e.g., InternVL). Finally, it provides a summary and outlook on the advances in the field of continual learning.

3.13 Continual, Decentralized Compositionality for Sustainable Artificial Intelligence

Vincenzo Lomonaco (University of Pisa, IT)

- License** © Creative Commons BY 4.0 International license
© Vincenzo Lomonaco
URL <https://www.vincenzolomonaco.com/collage-lab>

Sustainable Artificial Intelligence envisions a shift from centralized, resource-intensive models toward decentralized, reusable systems that are both environmentally efficient and socially inclusive. By distributing intelligence across networks rather than concentrating it in isolated silos, decentralization enables collective innovation, transparency, and equitable access to AI capabilities. This approach aligns ecological sustainability with democratization: reducing the carbon and computational footprint of AI while empowering diverse communities to contribute, adapt, and benefit from shared knowledge. Through continual adaptation and compositional design, AI can evolve as a green, open, and participatory infrastructure – towards the definition of a circular economy, building new AI solutions from existing ones.

3.14 Modularity for Continual Robot Learning

Jorge Mendez-Mendez (Stony Brook University, US)

License  Creative Commons BY 4.0 International license
© Jorge Mendez-Mendez

Standard monolithic models struggle with continual learning, as they store knowledge in an unstructured manner that is difficult to update with new information. This talk argues that for domains with well-understood structures, like robotics, modular solutions offer a significantly more powerful framework for lifelong learning. Task and motion planning (TAMP) decomposes problems both temporally and functionally, providing a concrete exploration of this concept. I present a lifelong learning algorithm for TAMP’s parameter samplers, showing how a robot can continually improve at household chores. I then conclude by discussing how these modular principles relate to the future of continual learning in the era of foundation models.

3.15 What does it mean to accumulate knowledge in lifelong (machine) learning?

Martin Mundt (Universität Bremen, DE)

License  Creative Commons BY 4.0 International license
© Martin Mundt

When we speak of lifelong or continual (machine learning), we often refer to the ability to accumulate knowledge over long periods of time from possibly changing non-stationary data or experiences. This seems to have implicitly resulted in a tendency to believe the ability to “add on” knowledge without forgetting is the main ingredient to success. In this talk, I argue that this is not necessarily realistic. To this end, I show two examples, tackling a controlled small scale and a large-scale real-world scenario. In the first, “adding on” knowledge is problematic, because confounders, i.e. spurious correlations, change over time due to partial observability – leading to contradictions and the necessity to revise knowledge. In the second example, a dataset spanning 250 years of books, changes in language and society lead to knowledge becoming outdated, obsolete, or possibly changing substantially. In an analysis of training LLMs on this kind of data over time and applying content moderation tools it becomes clear that our current (continual learning) tools are not yet set up to tackle such scenarios adequately.

3.16 CoPeP: Continual Pretraining for Protein Language Models

Darshan Patil (MILA – Montreal, CA)

License  Creative Commons BY 4.0 International license
© Darshan Patil

In recent years, protein language models (pLMs) have gained significant attention for their ability to capture the structure and function of proteins, accelerating the discovery of new therapeutic drugs. These models are typically trained on large, evolving corpora of proteins that are continuously updated by the biology community. motivating the need for continual

learning. As a result, we introduce the Continual Pretraining of Protein Language Models (CoPeP) benchmark, a novel benchmark for evaluating continual learning approaches on pLMs. We evaluate methods from the continual learning literature spanning replay, unlearning, and plasticity-based methods, some of which have never been applied to models and data of this scale. Our findings reveal that continual pretraining can improve the perplexity of the model over the traditional, non-continual training and several continual learning methods do outperform naive continual pretraining.

3.17 A Metacognition Lens on Continual Foundation Model Learning: Progress & Challenges

Ameya Prabhu (Universität Tübingen, DE)

License © Creative Commons BY 4.0 International license

© Ameya Prabhu

Joint work of Ameya Prabhu, Mathias Bethge

I broadly compare human metacognition with language models. Humans build a coherent map of the world and monitor their own thinking while “reading” and “writing” throughout their life. The key question: Can LLMs do the same – continually learn from mistakes, backtrack, correct, check, and plan? I conjecture there are three pillars: metacognitive knowledge (learning of facts, procedures, strategies as internal knowledge offline i.e. across large text corpora), metacognitive regulation (given a problem at test-time, trying to continually planning, monitoring, verification, and self-correction but online), and metacognitive experience (combining the both after each task – reflecting on past to continually improve future choices via world modeling). Open problems include how to self-improve with better data, algorithms, and pipelines; how to measure retention and transfer at the sample level; how to trade off facts, tool use, and strategy; how to detect and recover from errors using language feedback; how to evaluate and refine planning; and how to formalize continual world modeling while assessing benchmarks, metrics, and methods. The goal overall pitched is to chart progress and surface open problems across these three metacognitive axes in continual foundation-model learning.

3.18 Beyond Zero-Shot Generalization: Strengthening Vision-Language Models with Adaptation & Personalization

Elisa Ricci (University of Trento, IT)

License © Creative Commons BY 4.0 International license

© Elisa Ricci

Groundbreaking achievements in Vision-Language pretraining have increased the interest in crafting Vision-Language Models (VLMs) that can understand visual content alongside natural language, enabling a new definition of zero-shot classification. Despite huge pretraining databases, VLMs still face limitations, suffering from performance degradation in case of large train-test dissimilarity and requiring the design of highly generalizing textual templates. Test-Time Adaptation (TTA) and Few-Shot Adaptation (FSA) can effectively improve the robustness of VLMs by adapting a given model to online inputs or to a specific downstream

task given a few annotated examples. In this talk, I will present an overview of a few recent works from my research group on TTA, FSA and other strategies aimed at improving the robustness of VLMs across various downstream tasks.

3.19 Some reflections on Deep Continual Learning in the foundation model era

Tinne Tuytelaars (KU Leuven, BE)

License  Creative Commons BY 4.0 International license
© Tinne Tuytelaars

In this talk, I briefly discuss two recent works from our lab. The first one investigates whether it is possible to predict which samples will be forgotten when learning continually. We find that what is learned slowest is forgotten first. Adapting the sampling accordingly when constructing a replay buffer can reduce forgetting. In the second work, we investigate continual learning with access to all data from previous tasks, with the goal to make the training more compute efficient. We show that careful initialization, regularization and batch sampling can make training twice as fast with similar or better performance than joint incremental learning.

3.20 Do we need test-time adaptation for VLMs? The side-effects of unlearning

Bartłomiej Twardowski (Computer Vision Center – Barcelona, ES)

License  Creative Commons BY 4.0 International license
© Bartłomiej Twardowski

This presentation investigates the efficacy of test-time adaptation (TTA) for Vision Language Models (VLMs), particularly in more realistic scenarios. Our preliminary findings indicate a consistent degradation in VLM performance with increasing levels of visual corruption that cannot be overcome by improved prompting techniques. Furthermore, the study explores the unintended consequences of unlearning, demonstrating how the removal of a single concept can significantly impact the representation and behavior of other, seemingly unrelated concepts. Mechanistic Interpretability (MI) offers a promising framework for a more nuanced understanding of learning and unlearning processes within continual learning settings.

3.21 How to look at forgetting?

Gido van de Ven (University of Groningen, NL)

License © Creative Commons BY 4.0 International license
 © Gido van de Ven
Joint work of Timm Hess, Eli Verwimp, Gido M. van de Ven, Tinne Tuytelaars
Main reference Timm Hess, Eli Verwimp, Gido M. van de Ven, Tinne Tuytelaars: “Knowledge Accumulation in Continually Learned Representations and the Issue of Feature Forgetting”, *Trans. Mach. Learn. Res.*, Vol. 2024, 2024.
URL <https://openreview.net/forum?id=aHtZuZfHcf>

Continual learning research has shown that neural networks suffer from catastrophic forgetting “at the output level”, but it is debated whether this is also the case at the level of learned representations. We show that, even though forgetting in the representation (i.e., feature forgetting) can be small in absolute terms, when measuring relative to how much was learned during a task, forgetting in the representation tends to be just as catastrophic as forgetting at the output level. We also show that this feature forgetting is problematic as it substantially slows down the incremental learning of good general representations (i.e., knowledge accumulation).

3.22 The Butterfly Effect of Minimal Updates on Generative Visual Models

Joost van de Weijer (Computer Vision Center – Barcelona, ES)

License © Creative Commons BY 4.0 International license
 © Joost van de Weijer
Main reference Héctor Laria Mantecon, Alex Gomez-Villa, Imad Eddine Marouf, Kai Wang, Bogdan Raducanu, Joost van de Weijer: “Assessing Open-world Forgetting in Generative Image Model Customization”, *CoRR*, Vol. abs/2410.14159, 2024.
URL <https://doi.org/10.48550/ARXIV.2410.14159>

In this talk, I investigate forgetting in visual foundation models for image generation. We introduce the term open-world forgetting to refer to the fact that forgetting in foundation models can occur throughout its vast latent space. We look at the special case of minimal updates, which refer to tasks that only require small parameter updates to the model. We propose three methods to measure catastrophic forgetting in DMs, and show that for all three these methods we observe catastrophic forgetting even for the case of minimal model updates.

3.23 Continual Learning for Multi-modal Human-centric Applications

Liyuan Wang (Tsinghua University – Beijing, CN)

License © Creative Commons BY 4.0 International license
 © Liyuan Wang

Continual learning is a fundamental mechanism for enabling long-term adaptation and knowledge accumulation in intelligent systems. However, in the era of large-scale pretraining, it remains a critical challenge to extend continual learning to dynamic, heterogeneous real-world environments. This talk presents our recent efforts on continual learning in pretrained models, with a focus on two key directions. First, we propose a modality-heterogeneous

continual pretraining framework for multi-modal physiological signal generation, enabling real-time and robust monitoring of human health conditions. Second, inspired by the spatial cognition mechanisms of the biological brain, we develop embodied agents that construct and refine cognitive maps through continuous collection of spatial knowledge, thus equipping multi-modal language models with strong long-horizon generalization in complex environments. Together, these advances point toward the development of brain-inspired embodied intelligence with lifelong adaptability.

4 Working Groups

4.1 Working Group on Main Changes for Continual Learning in the Foundation Model Era

Participants identified several key ways in which continual learning in the foundation model era differs from classical settings. First, starting from large pretrained models changes the plasticity–stability trade-off: the initial model already encodes vast knowledge. Second, high-quality generative models enable synthetic data and self-improvement loops reminiscent of generative replay. In addition, parameter-efficient methods (e.g., LoRA) redefine how adaptation is performed. Third, evaluation becomes significantly harder: forgetting is less visible, benchmarks are costly and incomplete, and new metrics are needed. Finally, the scientific context has changed: scale and financial constraints limit experimentation within the academic context, emergent properties and in-context learning alter adaptation dynamics, and real-world scenarios increasingly motivate CL.

4.2 Working Group on Memory in Continual Learning in the Foundation Model Era

In this workgroup, the discussion was about how to optimally employ memory during continual learning of foundation models. Especially the newly introduced technique of prompt learning provides efficient ways for fast adaptation. The discussion stressed the parallels with memory in neuroscience: an agent receives a stream of data and must decide what to retain, compress, or forget over time. In the age of foundation and agentic models, memory extends beyond stored data to include the neural network parameters, external mechanisms like retrieval-augmented generation (RAG), and token-based context handled via attention, each solving complementary problems. Key challenges include handling long and redundant contexts, deciding which information truly matters, and updating memory without supervision, possibly via compression or separate memory-management models.

This topic has been chosen by the participants to aim to write a perspectives paper on in the near future.

4.3 Working Group on Forgetting in the Foundation Model Era

This working group focused on how to measure forgetting of foundation models during continual updates. Measuring catastrophic forgetting becomes especially difficult when updating foundation models due to the scale of their knowledge and the cost of evaluation.

In classical continual learning, forgetting is typically quantified by re-evaluating performance on previously seen tasks, but for foundation models this approach is often infeasible: comprehensive benchmarks are extremely expensive to run and only sparsely cover the breadth of capabilities such models encode. This motivates research on efficient evaluation strategies for continual adaptation of foundation models. Promising directions include reduced evaluation protocols that preserve broad capability coverage, as well as methods that infer forgetting from changes in model parameters or internal representations, enabling targeted assessment of vulnerable knowledge without requiring full-scale benchmark evaluation.

4.4 Working Group on Benchmarks, Setups, and Evaluation in the Foundation Model Era

Participants highlighted several open challenges around benchmarks, setups, and evaluation for continual learning with foundation models. A first issue is cost: evaluation can easily exceed fine-tuning itself, motivating a rethinking of benchmark design toward compute-efficient protocols, such as reduced task sequences, proxy or synthetic datasets. Second, while forgetting and adaptability remain core challenges, their relative importance is application-dependent, and different forms of forgetting (e.g., factual knowledge, reasoning ability, or core behaviors) may matter unequally; however, systematic large-scale evaluations comparing methods are still largely missing. Third, participants emphasized grounding continual learning in concrete applications – such as unlearning, personalization, deepfake detection, or tool use – to clarify objectives and expose real-world failure modes. Overall, the group argued that current benchmarks are often artificial and insufficient and new benchmarks need to be designed.

Participants

- Rahaf Aljundi
Toyota Motor Europe –
Zaventem, BE
- Andrew D. Bagdanov
University of Florence, IT
- Lucas Caccia
Microsoft Research –
Montréal, CA
- Antonio Carta
University of Pisa, IT
- Laurent Charlin
HEC Montréal, CA
- Jonghyun Choi
Seoul National University, KR
- Barbara Hammer
Universität Bielefeld, DE
- Tyler Hayes
Georgia Institute of Technology –
Atlanta, US
- Timm Felix Hess
KU Leuven, BE
- Christopher Kanan
University of Rochester, US
- Dhireesha Kudithipudi
University of Texas –
San Antonio, US
- Xialei Liu
Nankai University – Tianjin, CN
- Vincenzo Lomonaco
University of Pisa, IT
- Jorge Mendez-Mendez
Stony Brook University, US
- Martin Mundt
Universität Bremen, DE
- Darshan Patil
MILA – Montreal, CA
- Ameya Prabhu
Universität Tübingen, DE
- Elisa Ricci
University of Trento, IT
- Tinne Tuytelaars
KU Leuven, BE
- Bartłomiej Twardowski
Computer Vision Center –
Barcelona, ES
- Gido van de Ven
University of Groningen, NL
- Joost van de Weijer
Computer Vision Center –
Barcelona, ES
- Liyuan Wang
Tsinghua University –
Beijing, CN



Competitions and Empirical Evaluations in Automated Reasoning

Johannes K. Fichte^{*1}, Matti Järvisalo^{*2}, Aina Niemetz^{*3},
Andreas Niskanen^{†4}, and Guido Tack^{*5}

- 1 Linköping University, SE. johannes.fichte@liu.se
- 2 University of Helsinki, FI. matti.jarvisalo@helsinki.fi
- 3 Stanford University, US. niemetz@cs.stanford.edu
- 4 University of Helsinki, FI. andreas.niskanen@helsinki.fi
- 5 Monash University – Clayton, AU. guido.tack@monash.edu

Abstract

Solver competitions and practical problem solving challenges are a cornerstone in the field of Automated Reasoning (AR). They drive innovation by providing a platform for benchmarking, empirical evaluation, standardization of robust tools and methodologies, and identify challenges from research and industry. These events not only showcase the latest advancements in solver technology but also help to establish best practices for reliability and performance assessment. Organizing such competitions presents significant challenges, including the selection of representative benchmarks, the development of fair evaluation metrics, and ensuring result reproducibility. It is widely acknowledged that continued community engagement is essential for tackling these challenges and strengthening collaboration among organizers, developers, users, and reviewers. This report documents the program and the outcomes of Dagstuhl Seminar “Competitions and Empirical Evaluations in Automated Reasoning” (25441), which centered around competition challenges, discussed questions and solutions with the aim to build a community of practice of competition organization and empirical evaluation in AR.

Seminar October 26–31, 2025 – <https://www.dagstuhl.de/25441>

2012 ACM Subject Classification Computing methodologies → Symbolic and algebraic algorithms; Computing methodologies → Artificial intelligence; Mathematics of computing → Mathematical software performance; Software and its engineering → Empirical software validation; Computing methodologies → Knowledge representation and reasoning

Keywords and phrases automated reasoning, competitions, constraint solving, design of empirical experiments, empirical evaluation

Digital Object Identifier 10.4230/DagRep.15.10.135

1 Executive Summary

Johannes K. Fichte (Linköping University, SE)

Matti Järvisalo (University of Helsinki, FI)

Aina Niemetz (Stanford University, US)

Guido Tack (Monash University – Clayton, AU)

License © Creative Commons BY 4.0 International license
© Johannes K. Fichte, Matti Järvisalo, Aina Niemetz, and Guido Tack

Automated Reasoning (AR) emerged as a field in the late 1960s. Its sub-areas cover different aspects of deductive reasoning as practiced in mathematics and formal logic, such as automated theorem proving (ATP) and propositional satisfiability (SAT), among others.

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Competitions and Empirical Evaluations in Automated Reasoning, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 135–154
Editors: Johannes K. Fichte, Matti Järvisalo, Aina Niemetz, Andreas Niskanen, and Guido Tack



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Practical and theoretical research enabled ground-breaking success for vast applications of formal methods. At the core of this success are incredibly sophisticated and complex pieces of software, so-called *solvers*, which tackle specific problems in sub-areas of AR. Recurring solver competitions play a significant role in the practical success. They enable communities to showcase the current state-of-the-art, document practical advancements, identify challenges from research and industry, and push solver developers to not only aim for reliable and robust tools for a wide range of applications, but to improve their tools beyond current limitations.

Competitions often set standards when it comes to scientific empirical evaluations, with significant impact on the research community and their empirical methods. Organizing a competition that yields meaningful results both for developers and users of the tools is challenging – a tremendous amount of work. And, since competitions serve as a driving force for new advancements in the area, certain decisions related to competition organization have a significant impact as they often establish standards for formats and experimental evaluations. Thus, competition organizers, competition participants, authors, and reviewers all face similar issues. A central goal of this seminar was to establish a bridge between various competition organizers to enable and activate future collaborations.

Participants

The Dagstuhl Seminar brought together organizers and stakeholders from major AR-related competitions, namely,

- the CADE ATP System Competition (CASC) [26, 25],
- the SAT Competition (SAT) [8, 9],
- the MaxSAT Evaluation [3, 2],
- the Satisfiability Modulo Theories Competition (SMT-COMP) [21, 30],
- the Competition on Software Verification (SV-COMP) [4, 5],
- the MiniZinc Challenge [27, 24, 23],
- the Model Counting Competition [13, 11, 12],
- the Pseudo-Boolean Competition (PB) [22],
- the Hardware Model Checking Competition (HWMCC) [7, 6],
- the Termination Competition (termComp) [15, 17],
- the Reactive Synthesis Competition (SYNTCOMP) [18, 19],
- the Answer Set Programming (ASP) Challenge [16, 10],
- the LP/CP programming contest (Prolog, ASP, SAT, CLP) [1],
- the Competition on Computational Models of Argumentation (ICCMA) [29, 20], and
- the Planning Competition (IPC) [14, 28].

Program, Focus, and Discussions

The program was structured around (i) competition survey talks and challenge collection; (ii) tutorials into benchmark selection, infrastructure, and evaluation; (iii) panels and discussion blocks to surface shared challenges; and (iv) outcomes to define concrete follow-ups. The first two days started with mapping the competition landscape. Numerous short talks presented individual competitions and challenges to establish a shared baseline across solver communities and evaluation traditions. Tutorial talks throughout the week focused on benchmark selection (dataset choice, portfolios, robustness/diversity, anomaly detection) and on evaluation infrastructure, including widely used standardized benchmarking execution and cloud-based benchmarking systems. Evaluation-focused talks presented long-term efforts on benchmark databases, longitudinal benchmark analysis and libraries, cloud infrastructure,

and analytical approaches to competition results. A joint session with Research Meeting 25444 on Better Benchmarking Setups for Optimisation: Design, Curation and Long-Term Evolution provided insights into lessons from continuous optimization. Participants collected shared issues and targeted discussions to identify cross-cutting themes, for example, infrastructure, bias, repeatability, ranking methods, and interfaces. A dedicated panel on benchmarks and discussion sessions consolidated perspectives. Numerous discussion slots charted options for the “FLoC Olympics 2026”, participation/collaboration pathways, including spontaneous initiatives for contributing benchmark instances, and discussions on broader planning-competition perspectives.

Outcomes

Participants recognized that there is currently no universal methodology that guarantees robust, fair, and durable methods for competitions, benchmarking, and empirical evaluation in Automated Reasoning. Instead, participants converged on the view that good practice is necessarily context-dependent (competition goals, solver ecosystems, and community norms), requiring proper explanations and clear experimental design, while still benefiting from shared principles, shared resources, and a more precise articulation of trade-offs. Discussions emphasized, for example, the following aspects.

Benchmark sets should preferably be neutral, since this cannot always be guaranteed. Clearly highlighted pros and cons can be more valuable. Too much emphasis on the “best” result may limit development to what “progress” should look like, obstructing long-term research possibilities. Participants emphasized that a competition should include a well-argued benchmark selection, reduce overfitting to benchmark sets, enable representativeness beyond annual competition cycles (by reusing settings in paper submissions), facilitate new trends, and explore new applications.

Participants agreed that competitions are not purely technical; they serve as a tool for building communities. A joint infrastructure that simplifies coordination, research efforts, and stable evaluation would be highly appreciated. Participants argued that long-term discussions on standards for empirical methodology would be beneficial. Moreover, competition organizers should actively balance between a narrow leaderboard focusing on winners and detailed scientific reporting, including clear documentation of configurations, available resources, and constraints, different (possibly opposing) measures, and ranking approaches that reflect different needs. Participants favored explicit decisions that are transparent and understandable.

Moreover, practical considerations to simplify competition organization and enable easy, replicable evaluations, such as shared tooling, access to compute infrastructure (public HPC clusters and cloud resources), replication and artifact evaluation, and software quality, were discussed. What comes to infrastructure, developers and competition organizers often maintain their own infrastructure, as modern HPC research environments currently do not provide stable, reliable, and replicable execution for evaluating AR solvers. Therefore, the need to engage with HPC operators to establish a standard setup and run configurations on HPC environments was identified. In this context, the seminar discussed the future of StarExec, including perspectives on StarExec in the Cloud and the sustainability challenges of centralized evaluation platforms (governance, funding, maintenance, and risk mitigation).

Several coordination endeavors and community resources emerged spontaneously during the week. These included:

- an initiative on contributing benchmark instances (mysolvetimesout.org) aiming at lowering submission barriers and improving crediting practices (Daniel Le Berre);

- joining competitions focusing on onboarding and participation pathways (Laurent Simon);
- controversial evaluation statements and benchmarking (Ciran McCreesh)
- discussion of collaboration models to foster cross-competition exchange of tooling, benchmarks, and evaluation know-how (Marie Anastacio);
- distribution and dissemination of competition reports (Sophie Turret);
- designing variable but meaningful complementing rankings (Oliver Roussel);
- a joint website on AR competitions as a centralized entry point for competition information, best-practice guidance, and shared resources (Geoff Sutcliffe); and
- an initiative on requirements to establish a European competition infrastructure (Ciran McCreesh).

The participants agreed during the closing session to follow-up in particular on

- mechanisms for benchmark distribution and crediting;
- exploring publication venues and journal tracks for durable competition and evaluation outputs
- pursuing funding and coordination initiatives to support shared infrastructure;
- and planning future workshops, FLoC-related events, and an ERC Cost Action to maintain momentum and enable frequent meetings; and
- comprehensive survey paper synthesizing methodological lessons across competition ecosystems.

References

- 1 Mario Alviano. Lp/cp programming contest. <https://lpcp-contest.github.io/>, 2025.
- 2 Jeremias Berg, Matti Järvisalo, Ruben Martins, Andreas Niskanen, and Tobias Paxian. MaxSAT evaluation 2024 : Solver and benchmark descriptions. Technical report, Helsinki University Library, 2024.
- 3 Jeremias Berg, Matti Järvisalo, Ruben Martins, Andreas Niskanen, and Tobias Paxian. Maxsat evaluations: Evaluating the state of the art in maximum satisfiability solver technology. <https://maxsat-evaluations.github.io/>, 2025.
- 4 Dirk Beyer. Competition on software verification (sv-comp). <https://sv-comp.sosy-lab.org/>, 2025.
- 5 Dirk Beyer and Jan Strejček. Improvements in software verification and witness validation: SV-COMP 2025. In *Proceedings of the 31st International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS’25)*, pages 151–186, Hamilton, ON, Canada, 2025. Springer-Verlag.
- 6 Armin Biere, Nils Froleyks, and Mathias Preiner. Hardware model checking competition 2024. In *Proceedings of the 2024 Formal Methods in Computer-Aided Design (FMCAD’24)*, pages 1–1, 2024.
- 7 Armin Biere, Nils Froleyks, and Mathias Preiner. Hardware model checking competition. <https://hwccc.github.io/>, 2025.
- 8 Cayden Codel, Katalin Fazekas, Marijn Heule, and Ashlin Iser. SAT competition 2025. <https://satcompetition.github.io/2025/index.html>, 2025.
- 9 Cayden Codel, Katalin Fazekas, Marijn J. H. Heule, and Markus Iser. Proceedings of sat competition 2025 : Solver and benchmark descriptions. Technical report, TU Wien, 2025.
- 10 Carmine Dodaro, Christoph Redl, and Peter Schüller. The answer set programming challenge 2019. <https://sites.google.com/view/aspcomp2019/>, 2019.
- 11 Johannes K. Fichte and Markus Hecher. The model counting competitions 2021-2023. <https://arxiv.org/abs/2504.13842>, 2025.
- 12 Johannes K. Fichte, Markus Hecher, and Florim Hamiti. The model counting competition 2020. *ACM J. Exp. Algorithmics*, 26, oct 2021.

- 13 Johannes K. Fichte, Markus Hecher, and Arijit Shaw. Model counting competition. <https://mccompetition.org/>, 2025.
- 14 Daniel Fišer, Florian Pommerening, Jendrik Seipp, Javier Segovia-Aguas, Ayal Taitler, Scott Sanner, Joan Espasa Arxer, Enrico Scala, Ron Alford, Dominik Schreiber, and Gregor Behnke. International planning competition 2023. <https://ipc2023.github.io/>, 2023.
- 15 Florian Frohn, Jürgen Giesl, Georg Moser, Étienne Payet, Akihisa Yamada, and Dieter Hofbauer. Termination competition. https://termination-portal.org/wiki/Termination_Competition, 2025.
- 16 Martin Gebser, Marco Maratea, and Francesco Ricca. The seventh answer set programming competition: Design and results. *Theory and Practice of Logic Programming*, 20(2):176–204, 2020.
- 17 Jürgen Giesl, Albert Rubio, Christian Sternagel, Johannes Waldmann, and Akihisa Yamada. The termination and complexity competition. In Dirk Beyer, Marieke Huisman, Fabrice Kordon, and Bernhard Steffen, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, pages 156–166, Cham, 2019. Springer International Publishing.
- 18 Swen Jacobs, Guillermo A. Pérez, Remco Abraham, Veronique Bruyere, Michael Cadilhac, Maximilien Colange, Charly Delfosse, Tom van Dijk, Alexandre Duret-Lutz, Peter Faymonville, Bernd Finkbeiner, Ayrat Khalimov, Felix Klein, Michael Luttenberger, Klara Meyer, Thibaud Michaud, Adrien Pommellet, Florian Renkin, Philipp Schlehuber-Caissier, Mouhammad Sakr, Salomon Sickert, Gaetan Staquet, Clement Tamines, Leander Ten-trup, and Adam Walker. The reactive synthesis competition (SYNTCOMP): 2018-2021. <https://arxiv.org/abs/2206.00251>, 2024.
- 19 Swen Jacobs, Guillermo A. Pérez, and Philipp Schlehuber-Caissier. The reactive synthesis competition. <https://www.syntcomp.org/>, 2024.
- 20 Matti Järvisalo, Tuomo Lehtonen, and Andreas Niskanen. ICCMA 2023: 5th international competition on computational models of argumentation. *Artificial Intelligence*, 342:104311, 2025.
- 21 Martin Jonáš, François Bobot, David Déharbe, and Dominik Winterer. SMT-COMP 2025. <https://smt-comp.github.io/2025/>, 2025.
- 22 Olivier Roussel. Pseudo-Boolean competition 2025. <https://www.cril.univ-artois.fr/PB25/>, 2025.
- 23 Peter J. Stuckey, Ralph Becket, and Julien Fischer. Philosophy of the minizinc challenge. *Constraints*, 15(3):307–316, July 2010.
- 24 Peter J. Stuckey, Thibaut Feydy, Andreas Schutt, Guido Tack, and Julien Fischer. The MiniZinc challenge 2008–2013. *AI Magazine*, 35(2):55–60, Jun. 2014.
- 25 G. Sutcliffe. The 12th IJCAR Automated Theorem Proving System Competition - CASC-J12. *AI Communications*, 38(1):3–20, 2025.
- 26 Geoff Sutcliffe. The CADE ATP system competition. <https://tptp.org/CASC/>, 2025.
- 27 Guido Tack, Peter J. Stuckey, Jason Nguyen, Jip J. Dekker, Kevin Leo, and Maria Garcia de la Banda. The MiniZinc challenge. <https://www.minizinc.org/challenge/>, 2025.
- 28 Ayal Taitler, Ron Alford, Joan Espasa, Gregor Behnke, Daniel Fišer, Michael Gimelfarb, Florian Pommerening, Scott Sanner, Enrico Scala, Dominik Schreiber, Javier Segovia-Aguas, and Jendrik Seipp. The 2023 International Planning Competition. *AI Magazine*, 45(2):280–296, 2024.
- 29 Matthias Thimm, Johannes P. Wallner, Iosif Apostolakis, and Andrei Popescu. Ic-cma international competition on computational models of argumentation. <https://argumentationcompetition.org/>, 2025.
- 30 Tjark Weber, Sylvain Conchon, David Déharbe, Matthias Heizmann, Aina Niemetz, and Giles Reger. The SMT competition 2015-2018. *J. Satisf. Boolean Model. Comput.*, 11(1):221–259, 2019.

2 Contents

Executive Summary

Johannes K. Fichte, Matti Järvisalo, Aina Niemetz, and Guido Tack 135

Overview of Talks

ASP Competitions and LP/CP Programming Contests	
<i>Mario Alviano</i>	142
BenchCloud: A Platform for Scalable Performance Benchmarking	
<i>Dirk Beyer</i>	142
Improvements in Software Verification and Witness Validation: SV-COMP 2025	
<i>Dirk Beyer</i>	143
Termination and Complexity Competition	
<i>Florian Frohn</i>	144
Model Counting Competition 2020-2025: Evolution, Challenges, and Outlook	
<i>Markus Hecher</i>	144
Lessons Learned from the International Planning Competitions (Classical Setting)	
<i>Malte Helmert</i>	145
Global Benchmark Database	
<i>Ashlin Iser</i>	145
The History of SAT Competitions	
<i>Ashlin Iser</i>	146
SMT-COMP: International Satisfiability Modulo Theories Competition	
<i>Martin Jonáš</i>	146
MaxSAT Evaluations	
<i>Matti Järvisalo</i>	146
Intractable Decathlon: Benchmarking quantum hardware approaches on challenging discrete optimization problems	
<i>Thorsten Koch</i>	147
Using the Shapley Value to Analyze Competition Results	
<i>Lars Kotthoff</i>	147
Performance in the Context of Continuous Optimization Algorithms	
<i>Diederick Vermetten</i>	148
The MiniZinc Challenge	
<i>Jason Nguyen</i>	148
International Competitions on Computational Models of Argumentation (ICCMA)	
<i>Andreas Niskanen</i>	149
Hardware Model Checking Competition	
<i>Mathias Preiner</i>	149
SYNTCOMP: The Reactive Synthesis Competition	
<i>Guillermo A. Pérez</i>	149
Pseudo-Boolean Competitions	
<i>Olivier Roussel</i>	150

The RDDL Language and Planning Competitions: 15 Years of Trying to make Impact <i>Scott Sanner</i>	150
Understanding 20 Years of SMT Benchmarks <i>Hans-Jörg Schurr, Cesare Tinelli, Clark Barrett, François Bobot, Aina Niemetz, Pascal Fontaine, and Mathias Preiner</i>	150
Optimal Benchmark Selection via Instance Space Analysis <i>Kate Smith-Miles</i>	151
StarExec in Containers <i>Geoff Sutcliffe</i>	151
The CADE ATP System Competition <i>Geoff Sutcliffe</i>	152
Reliable Benchmarking: Requirements and Solutions <i>Philipp Wendler</i>	152
Panel Discussions	
Panel Discussion on “Benchmarks”	152
Participants	154

3 Overview of Talks

3.1 ASP Competitions and LP/CP Programming Contests

Mario Alviano (University of Calabria – Rende, IT)

License © Creative Commons BY 4.0 International license
© Mario Alviano

Main reference Martin Gebser, Marco Maratea, Francesco Ricca: “The Seventh Answer Set Programming Competition: Design and Results”, *Theory Pract. Log. Program.*, Vol. 20(2), pp. 176–204, 2020.

URL <https://doi.org/10.1017/S1471068419000061>

Declarative programming has long relied on competitions as catalysts for progress. The Answer Set Programming (ASP) Competitions emerged to provide a rigorous, comparable evaluation of ASP solvers, at a time when systems operated with incompatible syntaxes and no shared benchmark base. While the System Track focused on measuring solver performance, the Model & Solve Track encouraged creative modeling efforts across heterogeneous languages. The experience revealed both the value and the cost of human modeling: it pushed the community toward standardization (leading to the ASP-Core and ASP-Core-2 formats) and to the adoption of fully automated evaluation frameworks.

The LP/CP Programming Contest later revived the modeling challenge in a new, live, and cross-paradigm format combining ASP, Prolog, MiniZinc, and Picat, among other declarative paradigms. It reintroduced the human factor through short, competitive sessions emphasizing problem understanding, declarative thinking, and fun. Recent editions integrated visual checkers and ASP Chef to provide immediate, interpretable feedback, transforming evaluation into an interactive learning experience.

These initiatives show how competitions can evolve from testing machines to understanding people, revealing that progress in declarative reasoning depends as much on usability and modeling insight as on solver performance.

3.2 BenchCloud: A Platform for Scalable Performance Benchmarking

Dirk Beyer (LMU München, DE)

License © Creative Commons BY 4.0 International license
© Dirk Beyer

Joint work of Dirk Beyer, Po-Chun Chien, Marek Jankola

Main reference Dirk Beyer, Po-Chun Chien, Marek Jankola: “BenchCloud: A Platform for Scalable Performance Benchmarking”, in *Proc. of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 - November 1, 2024*, pp. 2386–2389, ACM, 2024.

URL <https://doi.org/10.1145/3691620.3695358>

Performance evaluation is a crucial method for assessing automated-reasoning tools. Evaluating automated tools requires rigorous benchmarking to accurately measure resource consumption, including time and memory, which are essential for understanding the tools’ capabilities. BenchExec, a widely used benchmarking framework, reliably measures resource usage for tools executed locally on a single node. This paper describes BenchCloud, a solution for elastic and scalable job distribution across hundreds of nodes, enabling large-scale experiments on distributed and heterogeneous computing environments. BenchCloud seamlessly integrates with BenchExec, allowing BenchExec to delegate the actual execution to BenchCloud. The system has been employed in several prominent international competitions in automated reasoning, including SMT-COMP, SV-COMP, and Test-Comp, underscoring

its importance in rigorous tool evaluation across various research domains. It helps to ensure both internal and external validity of the experimental results. This paper presents an overview of BenchCloud’s architecture and highlights its primary use cases in facilitating scalable benchmarking.

References

- 1 Dirk Beyer, Po-Chun Chien, and Marek Jankola. BenchCloud: A Platform for Scalable Performance Benchmarking. In Proc. ASE, pages 2386-2389, 2024. ACM. doi:10.1145/3691620.3695358

3.3 Improvements in Software Verification and Witness Validation: SV-COMP 2025

Dirk Beyer (LMU München, DE)

License © Creative Commons BY 4.0 International license
© Dirk Beyer

Joint work of Dirk Beyer, Jan Strejček

Main reference Dirk Beyer, Jan Strejček: “Improvements in Software Verification and Witness Validation: SV-COMP 2025”, in Proc. of the Tools and Algorithms for the Construction and Analysis of Systems - 31st International Conference, TACAS 2025, Held as Part of the International Joint Conferences on Theory and Practice of Software, ETAPS 2025, Hamilton, ON, Canada, May 3-8, 2025, Proceedings, Part III, Lecture Notes in Computer Science, Vol. 15698, pp. 151–186, Springer, 2025.

URL https://doi.org/10.1007/978-3-031-90660-2_9

The 14th edition of the Competition on Software Verification (SV-COMP 2025) evaluated 62 verification tools and 18 witness validation tools, making it the largest comparison of its kind so far. Out of these, 35 verification and 13 validation tools participated with an active support of teams led by 33 different representatives from 12 countries. The verification track of the competition was executed on a benchmark set of 33 353 verification tasks with C programs and 6 different specifications (reachability, memory safety, memory cleanup, overflows, termination, and data races) and 674 verification tasks with Java programs checked for assertion validity. Additionally, we considered 673 verification tasks with Java programs checked for runtime exceptions as a demo category. The validation track analyzed the witnesses generated in the verification track and newly also 103 handcrafted witnesses. To handle the increasing complexity of the competition, the organization committee has been established.

References

- 1 Dirk Beyer and Jan Strejček. Improvements in Software Verification and Witness Validation: SV-COMP 2025. In A. Gurfinkel and M. Heule, editors, Proceedings of the 31th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2025, Hamilton, Canada, May 3-8), part 3, LNCS 15698, pages 151-186, 2025. Springer.

3.4 Termination and Complexity Competition

Florian Frohn (RWTH Aachen, DE)

License  Creative Commons BY 4.0 International license
© Florian Frohn

The Termination and Complexity Competition (TermComp) focuses on automated termination and complexity analysis for various kinds of programming paradigms, including categories for term rewriting, integer transition systems, imperative programming, logic programming, and functional programming. In this talk, I given an overview of the competition, and I explain issues that we faced in recent years.

3.5 Model Counting Competition 2020-2025: Evolution, Challenges, and Outlook

Markus Hecher (University of Artois, CNRS – Lens, FR)

License  Creative Commons BY 4.0 International license
© Markus Hecher

Joint work of Arijit Shaw, Johannes K. Fichte, Markus Hecher

Main reference Johannes Klaus Fichte, Markus Hecher: “The Model Counting Competitions 2021-2023”, CoRR, Vol. abs/2504.13842, 2025.

URL <https://doi.org/10.48550/ARXIV.2504.13842>

The Model Counting Competition (MC) was conceived in 2020 with the goal of building a community around counting the number of satisfying assignments of propositional formulas, and of strengthening the diverse solving approaches for this task. At that time, the different facets of model counting and the variety of solving paradigms were still rather niche. Through the competition, these aspects were brought into focus, enabling systematic comparisons of strengths across approaches and fostering connections to different research communities, including applications and theoretical developments. This growing interest has since led to new approaches in proof formats and certified solving for counting, with applications even extending to quantum computing.

In the early editions, we primarily compared exact and approximate solvers side by side. Since then, the competition has evolved substantially, offering multiple tracks, evaluation measures, ranking schemes, and diverse benchmark sets. As a result, the complexity and effort required to organize each edition have steadily increased.

We identify ongoing challenges such as benchmark selection, evaluation environments, developer workload, benchmark publication pipelines, and the lack of a unified fuzzing and basic testing framework for participants. Furthermore, we discuss proposals for 2026 and beyond, focusing on how to strengthen the community and streamline future editions. Among the planned improvements are certificate outputs and numerical results in floating-point format (as opposed to fractions represented by numerator and denominator). We are confident that increased synergy with other competitions will bring substantial improvements and valuable insights for future editions.

3.6 Lessons Learned from the International Planning Competitions (Classical Setting)

Malte Helmert (Universität Basel, CH)

License © Creative Commons BY 4.0 International license

© Malte Helmert

URL <https://www.icaps-conference.org/competitions/>

The International Planning Competitions (IPC), hosted by the ICAPS conference series, have been a central part of automated planning research since the first competition in 1998. Since then, the IPC has branched out in multiple directions. In my talk, I discussed the “classical” branch, whose 9th edition took place in 2023.

One of the main challenges in running the IPC is devising fresh benchmark domains. The major difficulty in automated planning lies in adapting to unseen domains, and once a domain has been used and is publicly known, it is effectively burned for the competition purposes.

For its first ten years, the IPC was run in a very liberal fashion, with solver tuning happening alongside benchmark releases and no rigorous scoring framework. In 2008, the setup was tightened and has since remained largely stable, with smaller changes such as new competition tracks and the incremental evolution of the input format.

Besides describing the setup and history of the competition, the talk also covered the IPC’s stance on several controversial or challenging aspects of running competitions in automated reasoning, including scoring methods, how to handle incorrect solutions, and what to do about portfolios and planning systems that are only minor variants of solvers implemented by other teams.

3.7 Global Benchmark Database

Ashlin Iser (KIT – Karlsruher Institut für Technologie, DE)

License © Creative Commons BY 4.0 International license

© Ashlin Iser

Main reference Ashlin Iser, Christoph Jabs: “Global Benchmark Database”, in Proc. of the 27th International Conference on Theory and Applications of Satisfiability Testing, SAT 2024, Pune, India, August 21-24, 2024, LIPIcs, Vol. 305, pp. 18:1–18:10, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.

URL <https://doi.org/10.4230/LIPICS.SAT.2024.18>

Distributed management of instance metadata with GBD tools has not only been invaluable in organising the latest SAT competitions, but also for benchmarking and researching benchmarking methods. In my talk, I will provide an overview of GBD’s concepts and data model, and discuss recent developments. These include the efficient management of instance isomorphism classes and their integration into sustainable benchmarking tools. Finally, I will present a vision for the future of systematic, automated, data-driven algorithmic evaluations.

3.8 The History of SAT Competitions

Ashlin Iser (KIT – Karlsruher Institut für Technologie, DE)

License  Creative Commons BY 4.0 International license
© Ashlin Iser

Main reference Nils Froleyks, Marijn Heule, Ashlin Iser, Matti Järvisalo, Martin Suda: “SAT Competition 2020”, *Artif. Intell.*, Vol. 301, p. 103572, 2021.

URL <https://doi.org/10.1016/J.ARTINT.2021.103572>

The SAT competitions are a well-established series of international events that focus on the problem of propositional satisfiability. They showcase the latest algorithmic solutions to this central and challenging computer science problem, thereby encouraging further progress in the field. Organising these competitions is challenging in itself, with tasks including compiling a representative benchmark set, establishing sensible rules and conditions, and ensuring the necessary public infrastructure for benchmarking is in place. In this talk, I present the history of SAT competitions, highlight key milestones in their evolution, and discuss recent organisational developments. I will conclude with some discussion points to inspire future competitions.

3.9 SMT-COMP: International Satisfiability Modulo Theories Competition

Martin Jonáš (Masaryk University – Brno, CZ)

License  Creative Commons BY 4.0 International license
© Martin Jonáš

Joint work of Martin Jonáš, François Bobot, David Déharbe, Dominik Winterer

URL <https://smt-comp.github.io/>

The International Satisfiability Modulo Theories Competition (SMT-COMP) is an annual competition of state-of-the-art Satisfiability Modulo Theories (SMT) solvers. In this talk, I give details about the organization of the competition and the infrastructure that it uses. I also describe some issues that the competition faces now and potential issues that it will have to face in the future.

3.10 MaxSAT Evaluations

Matti Järvisalo (University of Helsinki, FI)

License  Creative Commons BY 4.0 International license
© Matti Järvisalo

Joint work of Matti Järvisalo, Fahiem Bacchus, Jeremias Berg, Ruben Martins, Andreas Niskanen, Tobias Paxian

URL <https://maxsat-evaluations.github.io>

We give a brief overview of the MaxSAT Evaluation series of competitive events.

3.11 Intractable Decathlon: Benchmarking quantum hardware approaches on challenging discrete optimization problems

Thorsten Koch (Zuse Institute Berlin, DE). Remote Talk.

License © Creative Commons BY 4.0 International license
© Thorsten Koch

Main reference Thorsten Koch, David E. Bernal Neira, Ying Chen, Giorgio Cortiana, Daniel J. Egger, Raoul Heese, Narendra N. Hegade, Alejandro Gomez Cadavid, Rhea Huang, Toshinari Itoko, Thomas Kleinert, Pedro Maciel Xavier, Naeimeh Mohseni, Jhon A. Montanez-Barrera, Koji Nakano, Giacomo Nannicini, Corey O'Meara, Justin Pauckert, Manuel Proissl, Anurag Ramesh, Maximilian Schicker, Noriaki Shimada, Mitsuharu Takeori, Victor Valls, David Van Bulck, Stefan Woerner, Christa Zoufal: “Quantum Optimization Benchmarking Library – The Intractable Decathlon”, CoRR, Vol. abs:2504.03832, 2025.

URL <https://arxiv.org/abs/2504.03832>

New hardware approaches are emerging, with quantum computers at the forefront, along with other systems such as data-flow machines, mem-computing, and bifurcation chips. All have in common their claim to “solve” challenging, i.e., NP-hard, combinatorial optimization problems more effectively than traditional methods. NP-hard problems are referred to as “intractable”, and, therefore, considered challenging. However, this characterizes the theoretical worst-case complexity for the entire class of problems. The assertion that finding a solution is exceedingly difficult pertains to the decision problem. In contrast, identifying some feasible solution to the optimization version of the problem is often straightforward. For instance, any permutation of cities constitutes a valid tour for the Traveling Salesperson Problem (TSP). The challenge lies in discovering the optimal tour and proving its optimality. The new approaches primarily can provide “good” solutions but fall short of proving optimality. The theoretical debate extends to whether and to what extent these problems can be approximated in polynomial time. However, the assurance an approximation algorithm offers is merely a lower bound on the solution’s quality. Numerous questions remain unanswered, and ultimately, the only method to evaluate the practical performance of heuristic algorithms is to benchmark them against relevant instances. We selected model-independent instances from a diverse set of ten different problem classes where classic exact and heuristic methods are known to have a difficult time. We will present the ten problem classes with some baseline results and some insights on their properties.

3.12 Using the Shapley Value to Analyze Competition Results

Lars Kotthoff (University of St Andrews, GB)

License © Creative Commons BY 4.0 International license
© Lars Kotthoff

Joint work of Lars Kotthoff, Alexandre Fréchette, Tomasz Michalak, Talal Rahwan, Holger H. Hoos, Kevin Leyton-Brown

Main reference Lars Kotthoff, Alexandre Fréchette, Tomasz Michalak, Talal Rahwan, Holger H. Hoos, Kevin Leyton-Brown: “Quantifying algorithmic improvements over time”, in Proc. of the 27th International Joint Conference on Artificial Intelligence, IJCAI’18, p. 5165–5171, AAAI Press, 2018.

URL <https://dl.acm.org/doi/10.5555/3304652.3304728>

It is surprisingly difficult to quantify an algorithm’s contribution to the state of the art. Reporting an algorithm’s standalone performance wrongly rewards near-clones while penalizing algorithms that have small but distinct areas of strength. Measuring an algorithm’s marginal contribution is better, but penalizes sets of strongly correlated algorithms, thereby obscuring situations in which it is essential to have at least one algorithm from such a set. Neither of these measures takes time into account, penalizing algorithms that are no longer state-of-the-art, but were when they were introduced.

In this talk, I will argue that contributions should be analyzed via a measure drawn from coalitional game theory, the Shapley value, and its time-sensitive cousin, the temporal Shapley value. The temporal Shapley Value maintains the desirable properties of the Shapley Value, but allows to take the time algorithms were introduced into account, within the context of the state of the art at the time. These measures characterize the contribution of an algorithm fairly and can yield insight into a research community's progress over time.

3.13 Performance in the Context of Continuous Optimization Algorithms

Diederick Vermetten (Sorbonne Université, CNRS, LIP6 – Paris, FR)

License  Creative Commons BY 4.0 International license
© Diederick Vermetten

When looking at an optimization algorithm, there is not a single way to judge its performance. Instead, the performance of an optimizer can be viewed as a three-dimensional shape, consisting of time (generally in terms of function evaluations), quality (either directly the function value, or relative to some known solution), and probability (as these algorithms are typically stochastic). By examining performance in this way, we can gain insight into how different algorithms complement each other. With this perspective in mind, we show some analysis examples, as well as various tools set up to facilitate benchmarking with flexible performance evaluation in mind.

3.14 The MiniZinc Challenge

Jason Nguyen (Monash University – Clayton, AU)

License  Creative Commons BY 4.0 International license
© Jason Nguyen

The MiniZinc Challenge is an annual solver competition in the Constraint Programming (CP) community held before the International Conference on Principles and Practice of Constraint Programming. In this presentation, we describe the motivation and goals of the Challenge, the format of the competition, and the infrastructure used to organize it. We also discuss issues faced when running the Challenge, as well as future directions the competition may move in the future.

3.15 International Competitions on Computational Models of Argumentation (ICCMA)

Andreas Niskanen (University of Helsinki, FI)

License © Creative Commons BY 4.0 International license
© Andreas Niskanen

Joint work of Matti Järvisalo, Tuomo Lehtonen, Andreas Niskanen

Main reference Matti Järvisalo, Tuomo Lehtonen, Andreas Niskanen: “ICCMA 2023: 5th International Competition on Computational Models of Argumentation”, *Artif. Intell.*, Vol. 342, p. 104311, 2025.

URL <https://doi.org/10.1016/J.ARTINT.2025.104311>

The series of International Competitions on Computational Models of Argumentation (ICCMA) aims at nurturing research and development of practical reasoning algorithms for computational models of argumentation. Organized biennially, ICCMA provides a snapshot of the current state of the art in algorithm implementations for central reasoning tasks in formal argumentation. In this talk, we provide an overview of ICCMA 2023, including a summary of competition tracks, details on various new developments introduced in this edition, the construction of benchmark sets, as well as considerations for future iterations of the competition.

3.16 Hardware Model Checking Competition

Mathias Preiner (Stanford University, US)

License © Creative Commons BY 4.0 International license
© Mathias Preiner

The Hardware Model Checking Competition is a series of competitive events to evaluate the state of the art of hardware model checking tools. Since its first edition in 2007, the competition took place 13 times. In this talk, I’ll give an overview on how the competition is organized.

3.17 SYNTCOMP: The Reactive Synthesis Competition

Guillermo A. Pérez (University of Antwerp, BE)

License © Creative Commons BY 4.0 International license
© Guillermo A. Pérez

Joint work of Guillermo A. Pérez, Swen Jacobs, Philipp Schlehuber-Caissier

Main reference Swen Jacobs, Guillermo A. Pérez, Remco Abraham, Véronique Bruyère, Michaël Cadilhac, Maximilien Colange, Charly Delfosse, Tom van Dijk, Alexandre Duret-Lutz, Peter Faymonville, Bernd Finkbeiner, Ayrat Khalimov, Felix Klein, Michael Luttenberger, Klara J. Meyer, Thibaud Michaud, Adrien Pommellet, Florian Renkin, Philipp Schlehuber-Caissier, Mouhammad Sakr, Salomon Sickert, Gaëtan Staquet, Clément Tamines, Leander Tentrup, Adam Walker: “The Reactive Synthesis Competition (SYNTCOMP): 2018-2021”, *Int. J. Softw. Tools Technol. Transf.*, Vol. 26(5), pp. 551–567, 2024.

URL <https://doi.org/10.1007/S10009-024-00754-1>

The Reactive Synthesis Competition (SYNTCOMP) is a competition for reactive synthesis tools. The competition’s goal is to collect benchmarks in a publicly available library and foster research in new tools for automatic synthesis of systems. SYNTCOMP is organized annually (since 2014) as a satellite event of CAV.

In this opportunity, I focused on conveying what is going well and what perceived issues there are.

3.18 Pseudo-Boolean Competitions

Olivier Roussel (University of Artois, CNRS – Lens, FR)

License  Creative Commons BY 4.0 International license
© Olivier Roussel

This talk is a short presentation of the competition of Pseudo-Boolean solvers. It presents the problem addressed by this competition, the goals of the competition, its general settings and the main issues in organizing it. The intent of these slides is to contribute to the reflexion on the organization of solvers competitions.

3.19 The RDDDL Language and Planning Competitions: 15 Years of Trying to make Impact

Scott Sanner (University of Toronto, CA)

License  Creative Commons BY 4.0 International license
© Scott Sanner

This talk will outline the history of the probabilistic planning competition and predecessor languages, the impetus for the creation of RDDDL and associated competitions to help the planning community have more impact, some rare successes, and the many challenges I've faced bridging RDDDL to various communities (Planning, Reinforcement Learning, and Operations Research) in achieving these goals.

3.20 Understanding 20 Years of SMT Benchmarks

Hans-Jörg Schurr (University of Iowa – Iowa City, US), Cesare Tinelli, Clark Barrett, François Bobot, Aina Niemetz (Stanford University, US), Pascal Fontaine, and Mathias Preiner (Stanford University, US)

License  Creative Commons BY 4.0 International license
© Hans-Jörg Schurr, Cesare Tinelli, Clark Barrett, François Bobot, Aina Niemetz, Pascal Fontaine, and Mathias Preiner

The SMT-LIB benchmark library is a large set of benchmarks for SMT solvers. It is used by the annual SMT competition to evaluate SMT solvers, and by researchers to study novel solving techniques. Effective use of the benchmark library often requires access to benchmark metadata, such as as the number of user defined symbols. We present a comprehensive metadata collection for the SMT-LIB benchmark library. It combines benchmark features with the results of all past SMT competition. Concretely, the catalog is implemented as a SQLite database. This allows users to use standard industry tools to perform queries, and the database to be distributed as a single file. In the future, the catalog will be distributed with the annual benchmark library release.

3.21 Optimal Benchmark Selection via Instance Space Analysis

Kate Smith-Miles (*The University of Melbourne, AU*)

License © Creative Commons BY 4.0 International license
© Kate Smith-Miles

Main reference Kate Smith-Miles, Mario Andrés Muñoz: “Instance Space Analysis for Algorithm Testing: Methodology and Software Tools”, *ACM Comput. Surv.*, Vol. 55(12), pp. 255:1–255:31, 2023.

URL <https://doi.org/10.1145/3572895>

This talk provided an overview of Instance Space Analysis [1] & the online tool MATILDA (matilda.unimelb.edu.au). The ISA methodology represents test instances in a 2-d mapping via a feature vector representation to visualise the relationship between instance features and algorithm performance. ISA enables i) the diversity and sufficiency of test instances to be established and augmented, and ii) the strengths and weaknesses of algorithms to be understood in terms of instance features. As such, ISA is a useful methodology for constructing suites of benchmark to rigorously “stress-test” algorithms under a wide variety of conditions represented by diverse test instances. A number of case studies were presented from combinatorial optimization (timetabling) & machine learning (ML). The library of existing instance spaces in MATILDA were shown, spanning various problems in optimization, machine learning & model fitting. Methods based on genetic algorithms were presented to show how an instance space can reveal target locations for evolving new instances to ensure a more comprehensive suite of benchmark instances to support competitions to produce diverse, unbiased and challenging benchmarks. Finally, using the instance space perspective to select an optimal subset of instances for competition benchmarks was discussed, building on published work using ML problems as case studies [2].

References

- 1 Smith-Miles, Kate and Muñoz, Mario Andrés. *Instance space analysis for algorithm testing: Methodology and software tools*. *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–31, 2023.
- 2 Pereira, João Luiz Junho and Smith-Miles, Kate and Muñoz, Mario Andrés and Lorena, Ana Carolina. *Optimal selection of benchmarking datasets for unbiased machine learning algorithm evaluation*. *Data Mining and Knowledge Discovery*, vol. 38, no. 2, pp. 461–500, 2024.

3.22 StarExec in Containers

Geoff Sutcliffe (*University of Miami, US*)

License © Creative Commons BY 4.0 International license
© Geoff Sutcliffe

Joint work of Geoff Sutcliffe, Andres Caicedo, David Fuenmayor, Jack McKeown

URL <https://tptp.org/Seminars/StarExecARC/OptionsDagstuhl/>

StarExec has been central to much progress in logic solvers over the last 10 years. StarExec Iowa has been decommissioned, and the smaller StarExec Miami is not able to support all the logic solver communities currently that used StarExec Iowa. In the long term StarExec will necessarily have to migrate to new compute environments. This talk describes work being done to reengineer StarExec using container technology. Supported by an Amazon Research Award, new versions of StarExec can be deployed in on your computer, on AWS, and your university cluster.

3.23 The CADE ATP System Competition

Geoff Sutcliffe (University of Miami, US)

License © Creative Commons BY 4.0 International license
© Geoff Sutcliffe

URL <https://tptp.org/Seminars/CASC/DagstuhlContents.html>

What is the CASC Problem? What is the CASC Solution? What is the CASC Issue?

3.24 Reliable Benchmarking: Requirements and Solutions

Philipp Wendler (LMU München, DE)

License © Creative Commons BY 4.0 International license
© Philipp Wendler

Joint work of Dirk Beyer, Stefan Löwe, Philipp Wendler

Main reference Dirk Beyer, Stefan Löwe, Philipp Wendler: “Reliable benchmarking: requirements and solutions”, Int. J. Softw. Tools Technol. Transf., Vol. 21(1), pp. 1–29, 2019.

URL <https://doi.org/10.1007/S10009-017-0469-Y>

Benchmarking is a widely used method in experimental computer science, in particular, for the comparative evaluation of tools and algorithms. As a consequence, a number of questions need to be answered in order to ensure proper benchmarking, resource measurement, and presentation of results, all of which is essential for researchers, tool developers, and users, as well as for tool competitions. We identify a set of requirements that are indispensable for reliable benchmarking and resource measurement of time and memory usage of automatic solvers, verifiers, and similar tools, and discuss limitations of existing methods and benchmarking tools. Fulfilling these requirements in a benchmarking framework can (on Linux systems) currently only be done by using the cgroup and namespace features of the kernel. We developed BenchExec, a ready-to-use, tool-independent, and open-source implementation of a benchmarking framework that fulfills all presented requirements, making reliable benchmarking and resource measurement easy. Our framework is able to work with a wide range of different tools, has proven its reliability and usefulness in the International Competition on Software Verification, and is used by several research groups worldwide to ensure reliable benchmarking.

4 Panel Discussions

4.1 Panel Discussion on “Benchmarks”

4.1.1 Panelists

- Erika Ábrahám (RWTH Aachen University, DE)
- Malte Helmert (University of Basel, CH)
- Mathias Preiner (Stanford University, US)
- Laurent Simon (Université de Bordeaux, FR)

4.1.2 Moderator

- Markus Hecher (CNRS, Artois University (CRIL), FR)

4.1.3 Discussion

The panelists identified that benchmarks, selection, and gathering are a shared bottleneck across AR competitions. They discussed intensively together with the seminar participants how to obtain benchmarks, what “good” or “bad” even means, and how benchmark choices steer research agendas. Panelists discussed that existing benchmark sets are often biased. For example, benchmarks are commonly drawn from where current tools already perform well or shaped by a small contributor base. These sets may not reflect real tasks, which is a recurring challenge in academia and industry. Sets may easily result in “instability” of competitions as different benchmark mixes yield different rankings. Sets that are not representative or are of different sizes and are used repetitively encourage overfitting from participating developers. This can turn competitions into narrow “examinations” rather than meaningful evaluations. What is it that we actually want to cover? What metrics, quality criteria, dimensions do we design? Do those reflect the initial intentions?

Variations such as in the planning competitions, where hard-to-ground instances have been included and featured, enable and drive new research agendas. The panelists repeatedly emphasised that benchmark curation needs to be considered a first-class research contribution. Communities should reward benchmark submissions, create artifact tracks, improve dissemination and outreach. The members of the panel suggested concrete ideas that worked well in existing competitions such as bring your own benchmarks, rotating benchmarks and selections to maintain diversity without excessive randomness, and compiling a representative benchmark set. The panelists argued that it might be more beneficial to add special tracks (e.g., learning/ML) rather than setting hard excluding requirements. In the past, this has been central to topics such as algorithm portfolios and the application of heavy algorithm selection tools in competitions. Soon, this might be a concern again when heuristics, parts of solvers, or entire solvers are produced by generative AI. However, the panelists also suggested that competition organizers and researchers conducting empirical evaluations should be explicit about the evaluation dimension (coverage of benchmarks/domains, difficulty, application relevance, and synthetic features). Reporting standards help to present a clear and honest picture where complete results including negatives and selection decisions are transparently revealed over the solver outperforms all previous solvers.

The discussion closed with a consensus that the community needs clearer empirical methodology for fair benchmark selection, better mechanisms for diverse and representative sets, and more inclusive access to evaluation infrastructure for teams with limited computational resources. Research questions that should be tackled by the community more intensively are: How to avoid running more or less the same benchmarks every year so that competitions become rather meaningless? How can we construct representative sets beyond annual competitions? How can we reduce overfitting of algorithms and solvers with effective measures that are reasonable for researchers and the community?

Participants

- Erika Ábrahám
RWTH Aachen University, DE
- Mario Alviano
University of Calabria –
Rende, IT
- Marie Anastacio
RWTH Aachen, DE
- Carlos Ansotegui
University of Lleida, ES
- Franz Baader
TU Dresden, DE
- Dirk Beyer
LMU München, DE
- Armin Biere
Universität Freiburg, DE
- Katalin Fazekas
TU Wien, AT
- Johannes Klaus Fichte
Linköping University, SE
- Florian Frohn
RWTH Aachen, DE
- Nils Froleyks
Johannes Kepler Universität
Linz, AT
- Markus Hecher
University of Artois, CNRS –
Lens, FR
- Keijo Heljanko
University of Helsinki, FI
- Malte Helmert
Universität Basel, CH
- Ashlin Iser
KIT – Karlsruher Institut für
Technologie, DE
- Matti Järvisalo
University of Helsinki, FI
- Martin Jonáš
Masaryk University – Brno, CZ
- Lars Kotthoff
University of St Andrews, GB
- Jean-Marie Lagniez
University of Artois, CNRS –
Lens, FR
- Daniel Le Berre
University of Artois, CNRS –
Lens, FR
- Ciaran McCreesh
University of Glasgow, GB
- Jason Nguyen
Monash University –
Clayton, AU
- Aina Niemetz
Stanford University, US
- Andreas Niskanen
University of Helsinki, FI
- Andy Oertel
Lund University, SE
- Guillermo A. Pérez
University of Antwerp, BE
- Mathias Preiner
Stanford University, US
- Le Quang Loc
University College London, GB
- Olivier Roussel
University of Artois, CNRS –
Lens, FR
- Simmo Saan
University of Tartu, EE
- Scott Sanner
University of Toronto, CA
- Dominik Schreiber
KIT – Karlsruher Institut für
Technologie, DE
- Hans-Jörg Schurr
University of Iowa –
Iowa City, US
- Thomas Sergeys
KU Leuven, BE
- Laurent Simon
University of Bordeaux, FR
- Kate Smith-Miles
The University of Melbourne, AU
- Geoff Sutcliffe
University of Miami, US
- Guido Tack
Monash University –
Clayton, AU
- Sophie Turret
INRIA – Villers-lès-Nancy, FR
- Philipp Wendler
LMU München, DE
- Akihisa Yamada
AIST – Tokyo, JP



Augmenting Human Creativity with AI

Tilman Dingler^{*1}, Elisa Giaccardi^{*2}, Steven Rick^{*3}, Irina Shklovski^{*4},
and Samangi Wadinambiarachchi^{†5}

1 TU Delft, NL. tilman.dingler@unimelb.edu.au

2 Polytechnic University of Milan, IT. elisa.giaccardi@polimi.it

3 NVIDIA – Santa Clara, US & MIT – Cambridge, US. srick@mit.edu

4 University of Copenhagen, DK. ias@di.ku.dk

5 The University of Melbourne, AU. samangi.w@unimelb.edu.au

Abstract

The advancements of available generative AI tools have the potential to change the creative processes of writers, artists, programmers, and designers. The Dagstuhl Seminar “Augmenting Human Creativity with AI” (25442) explored the grand challenges these tools pose to the discipline of design and human creativity more generally. Over the course of three days, academic researchers and Industry experts from the fields of Human-Computer Interaction, AI, and design came together to discuss how these technologies challenge accepted ideas of human creativity and what are the implications for the creative class when AI tools are used to augment creativity on a broader scale. The discussions revolved around three core themes: critical evaluation of AI-powered creative tasks, ethical implications of integrating AI into creative processes, focusing on ensuring the responsible and equitable use of these technologies, and the societal impact of delegating creative tasks to AI systems, including the potential consequences for the creative professions and cultural landscape. The seminar’s outcomes articulated a shared research agenda that moves beyond tool-building to focus on the design of infrastructure, practice, and educational frameworks that preserve creative agency and support diverse forms of expression. The seminar produced a collective vision, one that looks beyond building better tools and discusses an agenda towards cultivating the systems, norms, and learning environments needed to keep human creativity at the centre.

Seminar October 26–29, 2025 – <https://www.dagstuhl.de/25442>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Human-centered computing → Human computer interaction (HCI)

Keywords and phrases Design, Generative AI, Human Creativity, Human-AI Interaction

Digital Object Identifier 10.4230/DagRep.15.10.155

1 Executive Summary

Tilman Dingler (TU Delft, NL)

Elisa Giaccardi (Polytechnic University of Milan, IT)

Steven Rick (NVIDIA – Santa Clara, US & MIT – Cambridge, US)

Irina Shklovski (University of Copenhagen, DK)

License © Creative Commons BY 4.0 International license

© Tilman Dingler, Elisa Giaccardi, Steven Rick, and Irina Shklovski

The Dagstuhl Seminar “Augmenting Human Creativity with AI” brought together 27 international researchers, artists, and practitioners from design, art, computer science, and Human-Computer Interaction (HCI). The seminar examined how generative AI systems are

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Augmenting Human Creativity with AI, *Dagstuhl Reports*, Vol. 15, Issue 10, pp. 155–172

Editors: Tilman Dingler, Elisa Giaccardi, Steven Rick, Irina Shklovski, and Samangi Wadinambiarachchi



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

reshaping creative practice, not merely as productivity tools, but as infrastructures that redistribute agency, authorship, power, and responsibility. Throughout the three days filled with keynotes, provocations, discussions, and collaborative synthesis, the seminar explored three interrelated concerns: (1) how AI systems in creative practice should be critically evaluated and designed beyond efficiency and output quality; (2) how ethical governance, attribution, and ownership can be meaningfully embedded into creative AI infrastructures; and (3) how the increasing delegation of creative labor to AI systems reshapes cultural values, aesthetic norms, and societal power structures. A recurring tension throughout the seminar was that AI simultaneously amplifies and attenuates human creativity, which enables new forms of expression while constraining agency through opaque models, standardised interfaces, and platform dominance.

Two keynote talks framed these discussions. Baptiste Caramiaux's keynote, "Regaining Power Over AI: Insights from Artistic Practices," examined AI as a material, medium, tool, and companion, introducing a three-dimensional power framework (moral, structural, and cultural) to analyse how generative AI systems shape creative agency, authorship, and visibility. Maria Luce Lupetti's keynote, "Down the Algorithmic Rabbit Hole: Rethinking Design and Power in AI," challenged dominant narratives of AI "magic," highlighting how design practices can either obscure or expose the socio-technical assumptions embedded in AI systems. Together, these perspectives foregrounded creativity as a situated, contested, and political practice rather than purely technical capabilities. Building on this foundation, participants engaged in provocations and working sessions that culminated in five thematic working groups. These groups addressed: (1) the relationship between AI systems and the uniqueness of creative practices; (2) interaction paradigms beyond text-based prompting; (3) the evaluation of creativity beyond productivity metrics; (4) authorship, attribution, and disclosure in AI-mediated creation; and (5) AI literacy and education in creative domains. Rather than producing finalised solutions, the groups articulated research questions, design principles, and longer-term trajectories spanning master projects, PhD research, and large-scale collaborative initiatives.

Across the working groups, several cross-cutting research themes emerged. Participants emphasised a shift from large, general-purpose models toward smaller, situated, and personalized AI systems that better support craft, positionality, and local control. The re-authoring of creativity surfaced as a central concern, with debates around whether creators become curators of algorithmic outputs and how co-authorship with AI or simply use of AI-based tools might be made visible throughout the creative process. The seminar also highlighted the aesthetic and epistemic value of imperfection, arguing for AI systems that embrace noise, glitches, and discomfort rather than optimising for coherence and polish. In parallel, participants identified a significant evaluation gap between machine-learning benchmarks and human-centered notions of creativity, calling for process-oriented, experiential, and reflective evaluation methods. Finally, concerns about the invisibility of AI as infrastructure underscored the need to design for long-term critical reflection, even as AI systems become normalised and embedded in everyday creative practice. Overall, the Dagstuhl Seminar on Augmenting Human Creativity with AI provided an interdisciplinary forum for reframing creative AI use as a socio-technical, cultural, and ethical endeavour. The seminar's outcomes articulate a shared research agenda that moves beyond tool-building toward the design of infrastructures, practices, and educational frameworks that preserve agency, trigger reflection, and support diverse forms of creative expression. The insights in this report lay the groundwork for future collaborations seeking to shape creative AI systems that are not only powerful but also pluralistic, transparent, and human-centred.

2 Table of Contents

Executive Summary

Tilman Dingler, Elisa Giaccardi, Steven Rick, and Irina Shklovski 155

Overview of Talks

AI Models' Transparency Through Art: Between Enchantment and Critical Distance
Federico Bomba 159

Regaining Power Over AI: Insights from Artistic Practices
Bapiste Caramiaux 159

Training Otherwise: Sound Generation as a Speculative Site
Mona Hedayati 160

The Marginal, Overlooked, or Very Different GenAI
Lone Koefoed Hansen 160

Down the Algorithmic Rabbit Hole: Rethinking Design and Power in AI
Maria Luce Lupetti 161

From Prompting to Partnering: Designing AI Systems that Augment Human Creativity
Christoph Maerz 162

Against the Dark Force: Gravitational Attractors in GenAI
Dave Murray-Rust 162

Two Provocations: (1) AI commons, and (2) Small Data and Model Crafting.
Philippe Pasquier 163

Everyone Can Already Sing: Challenging Technosolutionism in Musicking (Provocation)
Courtney Reed 163

Enhancing Creativity in the Enterprise with AI
Justin D. Weisz 164

Working groups

How Can AI Systems Relate to the Uniqueness of Creative Practices?
Tilman Dingler, Lynda Hardman, Mona Hedayati, Steven Rick, Libuše Hannah Veprek, Boyu Xu, and Hiromu Yakura 165

Expanding Interaction Possibilities in AI Systems Beyond the Text Box
Elisa Giaccardi, Jesse Josua Benjamin, Bapiste Caramiaux, Sarah Fdili Alaoui, Janet Johnson, Christoph Maerz, Dave Murray-Rust, and Vasiliki Tsaknaki 166

Evaluation Metrics – How Do We Measure Creativity?
Tilman Dingler, Joseph Lindley, Philippe Pasquier, and Courtney Reed 166

Attributions
Irina Shklovski, Federico Bomba, Sarah Ciston, Laura Devendorf, Lone Koefoed Hansen, and Justin D. Weisz 167

AI Literacy and Education in Design and Creative Practices
Samangi Wadinambiarachchi, Tilman Dingler, Heekyoung Jung, and Maria Luce Lupetti 169

Performance

The Agents

Dave Murray-Rust, Joseph Lindley, Philippe Pasquier, and Courtney Reed 171

Participants 172

3 Overview of Talks

3.1 AI Models' Transparency Through Art: Between Enchantment and Critical Distance

Federico Bomba (Freie Universität Bozen, IT)

License  Creative Commons BY 4.0 International license
© Federico Bomba

Artificial Intelligence is often perceived as neutral and objective, masking biases, hidden labour, and infrastructural power. Transparency has become a pressing concern in generative AI, as these systems increasingly influence how knowledge, culture, and decisions are produced. This provocation presents *The Models*, an interactive installation by the artist duo *dmstfctn*, that stages large language models as improvised performers within the tradition of Italian *Commedia dell'Arte*. By embodying tendencies such as hallucination, antagonism, and servility, the work invites audiences to both believe in and question the machine. Building on this case, the talk introduces the notion of a *smark stance*: an oscillation between enchantment and critique enabled by a temporary suspension of disbelief. I argue that this ambivalent engagement allows generative AI to be experienced as neither infallible nor merely instrumental, opening a space where errors become legible and meaningful. Artistic practices like *The Models* suggest that fairness in AI can be cultivated through experiences that expose, rather than conceal, the systems' limits.

3.2 Regaining Power Over AI: Insights from Artistic Practices

Bapiste Caramiaux (Sorbonne University – Paris, FR)

License  Creative Commons BY 4.0 International license
© Bapiste Caramiaux

AI has made significant inroads into the creative and cultural sectors, both as a tool or medium and as a socio-political object. In this presentation, my aim is to highlight the friction that exists between technological innovation in AI and artistic practice. To do this, I review current knowledge on AI practices (pre-genAI and genAI) in artistic practices. I show that the latest generations of AI have diminished artists' agency over these technologies, both from a material and a cultural and political point of view. I then propose a framework for understanding the power relations between innovation and creative practices in three dimensions: cultural, structural and ethical. This framework is a way of encouraging more research into these socio-technical issues, beyond the opportunities and challenges that AI poses for the arts sectors.

3.3 Training Otherwise: Sound Generation as a Speculative Site

Mona Hedayati (University of Antwerp, BE & Concordia University, Montreal – CA)

License  Creative Commons BY 4.0 International license
© Mona Hedayati

(Artistic Provocation)

This artistic research demo presented an ongoing project that explores sound generation through a small, carefully curated dataset composed of socio-political sonic material. Countering the tendency to optimize for scale and efficiency that dominate contemporary machine learning practices, the project employs a small-scale sound corpus to train a generative sound model capable of producing unexpected and speculative sonic outputs. The results thus serve as an inquiry into how machine learning systems mediate complex situated lived realities. As such, the project foregrounds the ethics and aesthetics of working with data through embodied interaction with relevant material to construct a dataset, situating model training as a relational and iterative process unlike result-driven paradigms. In contrast to large-scale models reliant on massive amounts of compute and extractivist data logics, this work investigates how slow and small datasets and attentive training practices can open alternative modes of human-machine collaboration. By treating sound generation as a site for speculative experimentation beyond accurate representation, the project contributes to ongoing debates in artistic research and critical AI about the politics of data as knowledge, the materiality of training processes, and the possibilities of alternative machine learning practices.

3.4 The Marginal, Overlooked, or Very Different GenAI

Lone Koefoed Hansen (Aarhus University, DK)

License  Creative Commons BY 4.0 International license
© Lone Koefoed Hansen

Joint work of Søren Pold, Eva Eriksson, Lone Koefoed Hansen, Peter Dalsgaard

The purpose of my presentation was to focus on the marginalising aspects of technology adaptation as well as ways of using GenAI to suggest different kinds of histories, realities, and desires. I presented a recent essay that I co-authored [2] suggesting that applying a 10-year perspective makes it possible for us to see aspects of a technology, for instance what it marginalises, that are in many ways difficult to see on a day-to-day basis. In the essay, we use the example of how IoT technologies enables harm-doing in domestic abuse situations to ask how we might as researchers be better at understanding the potential harms that any technology (here: generative AI) will be able to lead to. I then proceeded to presenting two artists who use GenAI in their practice. First was the Swedish artist Arvida Byström, who in the piece “In the Clouds” (2024) worked with apps that can visually strip naked any human on a photograph. With this, she interestingly explores how the algorithms in the app “dream-up” impossible bodies, both expanding the idea of what a body might look like and exposing how the algorithm has a very narrow idea of the female body. I recently co-authored an article (in Danish) on this project [1].

Second artist is Danish artist Kristoffer Ørum, who uses GenAI to imagine what the cities and countries he works with could look like, had our social ideologies been different. Specifically, he asks what the city of Aarhus (Denmark) would look like if the positive energies

and ideologies of equality and creativity that we also find in hip-hop culture had been the dominant political organising principle. In total, the aim of my talk was to use my recent work/research to spark conversations on how we might think about which worlds are being built with GenAI.

References

- 1 Lone Koefoed Hansen, and Søren Bro Pold. *Kunstigt nøgen: Genforhandlet GenAI i Arvida Byströms In The Clouds*. (Danish) Passepartout-UdenTitel, 3:1-19, 2025.
- 2 Eva Eriksson, Lone Koefoed Hansen and Peter Dalsgaard. *We Won't Be Talking About GenAI in 2035, and That's a Problem*. Interactions, 32(6):26–30, 2025.

3.5 Down the Algorithmic Rabbit Hole: Rethinking Design and Power in AI

Maria Luce Lupetti (Polytechnic University of Torino, IT)

License  Creative Commons BY 4.0 International license
© Maria Luce Lupetti

This talk, Down the Algorithmic Rabbit Hole: Rethinking Design and Power in AI, explores how design practices shape, sustain, and can contest dominant narratives of artificial intelligence. Drawing on Lewis Carroll's Alice in Wonderland as a metaphorical framework, the presentation guides the audience through the multiple “characters” of AI – each symbolizing a tension in contemporary sociotechnical systems. The White Rabbit represents the lure of AI hype and techno-solutionism; the Queen of Hearts embodies irrational power and moral hypocrisy embedded in algorithmic norms; while the Cheshire Cat stands for ambiguity, multiplicity, and alternative imaginaries. Through a series of design research projects – including Shybo, Steering Stories, Dramatic Things, and more – the talk illustrates how speculative and participatory design methods can reveal and challenge hidden assumptions in AI narratives, from automation myths to biased data infrastructures. These projects reimagine AI not as a neutral or magical force but as a contested terrain of politics, labor, and meaning-making. The second half of the presentation introduces Participatory Design Justice for Ethical AI Transitions (PARJAI), a €1.3M research project funded by the Italian Ministry of Education and Research. PARJAI investigates how participatory design can advance social justice in AI transitions across domains such as health, work, and energy – addressing power by exclusion and opacity. Ultimately, the talk argues for a shift from designing for AI systems toward designing with diverse publics, acknowledging participation as both a site of empowerment and exploitation. By embracing ambiguity, plurality, and collective imagination, designers can move beyond enchantment toward more reflexive, just, and accountable AI futures.

3.6 From Prompting to Partnering: Designing AI Systems that Augment Human Creativity

Christoph Maerz (DFKI – Kaiserslautern, DE)

License  Creative Commons BY 4.0 International license
© Christoph Maerz

Recent advances in generative AI have shifted human–AI interaction from isolated prompting toward more sustained and integrated forms of collaboration. This provocation explores how this shift changes the role of AI systems in creative and knowledge-intensive work, moving from tools that produce outputs toward systems that actively participate in human sensemaking and ideation. Drawing on the theory of extended cognition, the talk examines how contemporary approaches such as AI assistants, deep research systems, multimodal avatars, and agent-based workflows can be embedded into ongoing creative processes. Examples from creative practice, education, and applied research illustrate how these systems influence the structure, pace, and framing of creative work. The provocation highlights both the potential and the open questions that arise when AI becomes a persistent partner in human creativity, inviting further discussion on how such systems should be designed, evaluated, and situated in real-world contexts

3.7 Against the Dark Force: Gravitational Attractors in GenAI

Dave Murray-Rust (TU Delft, NL)

License  Creative Commons BY 4.0 International license
© Dave Murray-Rust

In this talk, I develop the metaphor of black holes as a way to conceptualise the gravitational pull of generative AI systems on human culture and thinking. Gravity gives an invisible structure that spreads throughout space, shaping movement, modulating energy. As humans, we spend our time in negotiation with gravity, to stand up, to move around. We are aware of it, yet there are no single organs of gravity with which we sense it – we construct our understanding from a composite overview of proprioceptions and pressures. While we may fight against gravity, we also need it – without this invisible structure, there would be no fire, no rain, no potential for soaring and flying – simply floating. Gravity is a self reinforcing phenomenon: as masses are drawn together, their combined gravities increase until, in the limit, black holes form, structures in space time from which no light or other information can escape. We can read current AI development through this lens: the gravity of large models pulling our creative endeavour into particular directions and forms. Within this framing, gravitational attractors – black holes – emerge, that warp the cultural fabric in particular directions. Three of them are explored in the talk. Firstly, the idea that everything of interest can be coded and encoded in a text box. This creates a focus on written language, which is powerful, for its range and breadth, but elides all possibility of rhythm, texture, gesture and so on. Secondly, the functional form of transforming one thing into another: an interaction paradigm that gives interesting results when the subjects are rich, as the world of style transfer, or the transformation of text descriptions into images shows, but which misses other conceptual configurations for making use of technological intervention. Finally, the 'make it shiny' button, that transforms rough thoughts into polished text, automates musical mastering or fills in details in sketches and photographs. Here, not only is the

kind of perfection provided problematic through its bland centrality, it focuses the idea that perfection is the goal of creative endeavour, rather than the interesting things that happen along the way. In response to these black holes – which are only three among many – I propose useful creative structures for moving against gravity: intermediate level toolkits that allow constructive experimentation; development of new conceptual maps that highlight the space between the attractors; and direct attempts to break the programming that leads to the same endpoints.

3.8 Two Provocations: (1) AI commons, and (2) Small Data and Model Crafting.

Philippe Pasquier (Simon Fraser University – Surrey, CA)

License © Creative Commons BY 4.0 International license
 © Philippe Pasquier
URL <https://www.metacreation.net/>

From Minitel's hum to GPUs' glow,
 Our world computes, we reap, we sow.
 AI now hums beneath the art,
 A hidden hand in every part.
 Who owns this AI that learns our way?
 Who shapes the dreams machines convey?
 The Commons call for code set free,
 For artists' rights and equity.
 Not Big Data's greedy sprawl,
 Our hand-crafted models are small. Where care and craft replace the race,
 And creative AI finds its place.

3.9 Everyone Can Already Sing: Challenging Technosolutionism in Musicking (Provocation)

Courtney Reed (Loughbrough University London, GB)

License © Creative Commons BY 4.0 International license
 © Courtney Reed
Main reference Courtney N. Reed: "Enculturation and Value Encoding in the Design of Vocal DMIs", in Proc. of the CHIME Annual Conference 2024, CHIME '24, 2024.
URL https://www.courtneyreed.com/assets/pdf/Reed_CHIME24_VocalEnculturation.pdf
Main reference Rachel Freire, Courtney N. Reed: "Body Lutherie: Co-Designing a Wearable for Vocal Performance with a Changing Body", in Proc. of the International Conference on New Interfaces for Musical Expression, pp. 117–126, 2024.
URL <http://dx.doi.org/10.5281/zenodo.13904800>

(Artist Provocation)

The topic of the talk centred around a paper entitled "Anyone Can Sing," (the citation of which is intentionally not included to avoid drawing negative attention to the article, as that's not the purpose), which is a typical example of the technosolutionist mindset in the development of digital creative tools.

The paper discusses a generative system that allows a user to input spoken voice samples to generate a realistic output of their cleaned and edited singing voice in different styles and contexts. While technically sound and demonstrating cutting edge audio analysis and

style transfer technique, the applications demonstrate a social positioning that is embedded in technology: correcting typical singing voices to trends set in music consumption and, instead of singing “Happy Birthday” to your friend or child in your own voice (with your own identity and intention), to produce a cleaned and polished version of “you” as if you were a pop singer.

This work, and other similar papers, positions particular styles, such as classical Western traditions, as being superior to others, diminishes the value of “untrained” singing. This view is derived from colonial and capitalist values that position different (e.g., indigenous, non-white, untrained in the styles valued by upper class European society) vocal qualities as being more aesthetically appealing or “correct”, although there is no factual grounding in this typically-held belief.

This provocation centred the sociocultural biases underlying these digital technosolutionist approaches and questioned the deeply held beliefs that we have in society about music making and who is allowed to do it. In centring “bad” art as having inherent value and something deeply communicative between and about human beings, we can challenge these assumed norms and views of value in the creation of digital tools.

3.10 Enhancing Creativity in the Enterprise with AI

Justin D. Weisz (IBM TJ Watson Research Center – Yorktown Heights, US)

License © Creative Commons BY 4.0 International license
© Justin D. Weisz

Joint work of Justin D. Weisz, Kristina Brimijoin, Jule Ebeling, Morgan Foreman, Gabriel Gonzalez, Jessica He, Stephanie Houde, Siya Kunde, Michael Muller, Steven Ross, Dario Silva Moran

Ideation – the process of coming up with new ideas in response to a prompt or provocation – is a common activity in design thinking practice. Generative AI is quite capable of producing ideas in response to a prompt, but are they valuable? At IBM, we have created multiple tools to support ideation in groups: Koala, a Slack bot that participates in group brainstorming sessions, and Collaborative Canvas, a virtual canvas where people can use AI to generate ideas on virtual sticky notes, either publicly or privately. Studies of these tools have shown that participants find them to be valuable and participants preferred having AI participate in their group. Further, the best ideas were those where both humans and AI made some form of contribution. However, the presence of AI-generated ideas tended to reduce the overall diversity of ideas produced by the group, showing a nuanced effect of AI on the ideation process and outcomes.

References

- 1 Gonzalez, G. E., Moran, D. A. S., Houde, S., He, J., Ross, S. I., Muller, M. J., Kunde, S., & Weisz, J. D. (2024, March). Collaborative Canvas: A Tool for Exploring LLM Use in Group Ideation Tasks. In Joint Proceedings of the ACM IUI 2024 Workshops, co-located with the 29th Annual ACM Conference on Intelligent User Interfaces (IUI 2024), Greenville, South Carolina, USA, March 18, 2024, CEUR-WS.org/Vol-3660.
- 2 Ebeling, J. (2025, July). AI as a Creative Partner: Investigating how Different Levels of AI Proactivity Shape User Experience and Creative Output in Collaborative Idea Generation. Master’s Thesis, Bauhaus-Universität Weimar, Germany.
- 3 He, J., Houde, S., Gonzalez, G. E., Silva Moran, D. A., Ross, S. I., Muller, M., & Weisz, J. D. (2024, June). AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (pp. 1-14).

- 4 Houde, S., Brimijoin, K., Muller, M., Ross, S. I., Silva Moran, D. A., Gonzalez, G. E., Kunde, S., Foreman, M. A., & Weisz, J. D. (2025, March). Controlling AI Agent Participation in Group Conversations: A Human-Centered Approach. In Proceedings of the 30th International Conference on Intelligent User Interfaces (pp. 390-408).
- 5 Muller, M., Houde, S., Gonzalez, G., Brimijoin, K., Ross, S. I., Moran, D. A. S., & Weisz, J. D. (2024, June). Group Brainstorming with an AI Agent: Creating and Selecting Ideas. In International Conference on Computational Creativity.

4 Working groups

4.1 How Can AI Systems Relate to the Uniqueness of Creative Practices?

Tilman Dingler (TU Delft, NL), Lynda Hardman (CWI – Amsterdam, NL), Mona Hedayati (University of Antwerp, BE & Concordia University, Montreal – CA), Steven Rick (NVIDIA – Santa Clara, US & MIT – Cambridge, US), Libuše Hannah Veprek (Universität Tübingen, DE), Boyu Xu (Utrecht University, NL), and Hiromu Yakura (MPI für Bildungsforschung – Berlin, DE)

License © Creative Commons BY 4.0 International license

© Tilman Dingler, Lynda Hardman, Mona Hedayati, Steven Rick, Libuše Hannah Veprek, Boyu Xu, and Hiromu Yakura

Group 1 examined how AI systems can meaningfully engage with the situated, idiosyncratic, and often non-linear nature of creative practices. Participants identified a central limitation of contemporary generative AI tools: their tendency to promote design fixation, homogenization, and “echo chambers” through simplified input–output interaction models. Such paradigms obscure the rich, iterative, and exploratory processes that characterise creative work. In response, the group explored how AI systems might be designed to surface and support the in-between, i.e., the intermediate states, transformations, and uncertainties that occur between prompt and output. Emphasis was placed on interaction paradigms that invite tinkering, transparency, and serendipitous discovery, enabling creators to navigate, manipulate, and reflect on generative processes rather than merely consume results. The group proposed a multi-stage research agenda. First, ethnographic studies with creative practitioners would document existing workflows, values, and frictions in AI-assisted creation. These insights would inform a shared knowledge base mapping current AI tools, their affordances, and, critically, what is missing. Building on this foundation, the group envisioned speculative and experimental design explorations that challenge linear interaction models, for example, through visualizing generative processes, enabling non-linear intervention, or introducing productive friction. The overarching goal is to reconceptualise AI systems not as automated generators, but as creative partners that preserve authorship, support exploration, and amplify the uniqueness of individual creative voices.

4.2 Expanding Interaction Possibilities in AI Systems Beyond the Text Box

Elisa Giaccardi (Polytechnic University of Milan, IT), Jesse Josua Benjamin (TU Eindhoven, NL), Baptiste Caramiaux (Sorbonne University – Paris, FR), Sarah Fdili Alaoui (University of the Arts – London, GB), Janet Johnson (University of Michigan – Ann Arbor, US), Christoph Maerz (DFKI – Kaiserslautern, DE), Dave Murray-Rust (TU Delft, NL), and Vasiliki Tsaknaki (IT University of Copenhagen, DK)

License © Creative Commons BY 4.0 International license
 © Elisa Giaccardi, Jesse Josua Benjamin, Baptiste Caramiaux, Sarah Fdili Alaoui, Janet Johnson, Christoph Maerz, Dave Murray-Rust, and Vasiliki Tsaknaki

Group 2 explored how AI systems might better support “humaning” by expanding interaction beyond text-based interfaces toward multimodal, embodied, and relational forms of engagement. Rather than treating multimodality as a purely technical extension of existing interfaces, the group focused on what the generative and fuzzy qualities of contemporary AI introduce as fundamentally new interaction possibilities. Central to this exploration was a shift from efficiency-driven conversational models toward slower, more reflexive, and contemplative modes of interaction, including tactile, contextual, and sensual interfaces that engage the whole body.

The group also investigated “black box” strategies not as mechanisms for transparency, but as sites for intimacy, viscosity, and negotiated understanding, asking what it might mean to relate to AI without fully opening it or optimizing it. They discussed hegemonic narratives embedded in conversational AI, proposing alternatives such as conversational consent, bidirectional negotiation, and embodied manifestations of AI from products to robotics. They also foregrounded the importance of limits, articulating the concept of “un-AI”: negative spaces, black holes, and contexts where AI should deliberately not exist.

4.3 Evaluation Metrics – How Do We Measure Creativity?

Tilman Dingler (TU Delft, NL), Joseph Lindley (Lancaster University, GB), Philippe Pasquier (Simon Fraser University – Surrey, CA), and Courtney Reed (Loughborough University London, GB)

License © Creative Commons BY 4.0 International license
 © Tilman Dingler, Joseph Lindley, Philippe Pasquier, and Courtney Reed

Group 3 addressed the fundamental challenge of evaluating creativity in human–AI collaboration, arguing that prevailing evaluation approaches are ill-suited to creative practice. Current methods are often derived from machine learning benchmarks or productivity-oriented software metrics, and hence tend to reduce creativity to binary distinctions (e.g., good vs. bad, human vs. machine) or artifact-centric measures that neglect process, experience, and meaning. The group proposed a shift toward holistic, process-oriented evaluation frameworks that reflect creativity as situated, contested, and value-laden. Rather than privileging output quality alone, evaluation should account for dimensions such as expressivity, authorship, control, ownership, novelty, authenticity, trust, and the productive role of randomness and imperfection. Participants emphasised that creativity often emerges through struggle, iteration, and discomfort, which are qualities that are often obscured by efficiency-driven metrics. To operationalise this perspective, the group outlined several directions for future work. These

include developing functional applications that embed alternative evaluation logics, exploring participatory and narrative-based assessment methods, and supporting community-driven governance frameworks for evaluating creative AI systems. At a deeper level, the group called for a craft-based epistemology of AI, one that recognises tacit knowledge, embodied practice, and experiential judgment as legitimate forms of evaluation. Ultimately, the group aims to establish new standards for assessing creativity that resist reductive quantification and instead honour the complexity of creative work in AI-mediated contexts.

4.4 Attributions

Irina Shklovski (University of Copenhagen, DK), Federico Bomba (Freie Universität Bozen, IT), Sarah Ciston (Center for Advanced Internet Studies – Bochum, DE), Laura Devendorf (University of Colorado – Boulder, US), Lone Koefoed Hansen (Aarhus University, DK), and Justin D. Weisz (IBM TJ Watson Research Center – Yorktown Heights, US)

License © Creative Commons BY 4.0 International license

© Irina Shklovski, Federico Bomba, Sarah Ciston, Laura Devendorf, Lone Koefoed Hansen, and Justin D. Weisz

Group 4 explored the question of attribution as an important concept in contemporary creative practice delineating reception and authorship of creative products in the context of the use of Generative AI. New technologies often challenge traditional approaches to notions of authorship in creative production. As Benjamin argued, new technologies can transform what constitutes a creative product, challenging traditional notions of authenticity, uniqueness and value of creative work [1]. The emergence of generative AI systems has introduced another such challenge to our notions of creativity and authorship. In 2022, James M. Allen created a visual art piece entitled *Théâtre D’opéra Spatial* using Midjourney, which won a fine arts competition with a \$300 cash prize [4]. When the artist submitted a request to the United States Copyright Office for registration of the image, the request was refused because “the Work contains more than a *minimis* amount of AI-generated content” which can not be separated from human contribution and thus could not be said to be the work of “human authorship” [6]. As generative AI becomes ever more integrated into everyday creative practices, from essay writing to image creation, the question of attribution of authorship becomes ever more important across domains.

What does authorship look like when agentic AI systems are involved in co-production of content? Group and collective authorship of creative content has been a point of debate and concern for decades. Authorship and attribution are a type of contested performativity, a form of cultural production and form the basis for valuations of creative output [3]. What happens when we begin to think in terms of LLMs though – who is then speaking through the LLMs as these systems ingest human-generated content and ideas and then reproduce, remix, reformulate them in response to prompts? Who is included and who is excluded a culturally contested space of creativity? The practice of naming and attribution of the creative product can privilege the loudest voices and has repeatedly motivated explorations of alternatives to traditional naming and authorship expressions [2]. With the use of generative AI, how do we mark and signal authorship and what cultural changes might come about as “with the use of AI” becomes a common signal?

Past work at IBM on attribution for working with AI systems has explored attribution as a form of ownership or acknowledgement of contributions of different types made to a creative work. Acknowledging that different communities have different attribution standards

(the credit taxonomy), He, Houde and Wiesz [5] developed the IBM AI attribution toolkit (<https://aiattribution.github.io/>), providing a language to describe how AI was used to create a work. Clearly the new vocabulary is needed to signal and debate emerging notions of (co)authorship, ownership, credit, provenance when creative work is produced using generative AI systems. The questions become ever more complex however, as the proliferation of these systems across creative digital tools can result in creative practice that may be unaware of the use of Generative AI.

The group discussed many aspects of attribution, coming up with a loose taxonomy of what might constitute appropriate tracking and provenance marking:

- provenance, authorship, ownership, attribution
- origin or location – effort that goes into creation of output
- tracking from output to original – tracing back to original sources

The resulting discussion noted that in the end we arrive at impossible tasks of provenance tracking, such as: what would it take to trace the origin for every text? Creative content production is ubiquitous, but where are the limits to attribution? Does everything need to be attributed or only things generated with particular types of generative AI systems? Is the question of attribution rooted in the use of prompts or should we consider all generative AI tool use? Current debates propose technical marking of AI generated content through the C2PA provenance marking standards (<https://c2pa.org/>) or watermarking and other computationally more intensive approaches. Yet the question remains on what ought to be marked, when and why. What effects visibly or invisibly marking AI-generated or AI-assisted content will have on cultural production, creative pursuits and what comes to constitute authenticity for the work of art in the age of Generative AI. The group identified productive questions for research that is crucial for both tracking the normative evolution of how people will interpret AI-produced or AI-augmented content and, potentially, developing mechanisms for affecting this normative evolution by providing new mechanisms for marking and reflection.

References

- 1 Walter Benjamin. *The work of art in the age of mechanical reproduction*. Visual culture: Experiences in visual culture, vol. 4, pp. 114–37, 2006.
- 2 Erik D. Demaine and Martin L. Demaine. *Every Author as First Author*. arXiv:2304.01393 [cs.DL], 2023. Available at: <https://arxiv.org/abs/2304.01393>
- 3 Colin M. Fisher, Janice Sanchez, James Berry, and Wen-Xin Xie. *Who (and how many) made this? How crediting authorship affects creativity evaluations*. *Creative Success in Teams*, Elsevier, pp. 167–188, 2021. doi: <https://doi.org/10.1016/B978-0-12-819993-0.00009-6>
- 4 Matthew Gault. *An AI-Generated Artwork Won First Place at a State Fair Fine Arts Competition, and Artists Are Pissed*. Vice (via Way-back Machine), 2022. <https://web.archive.org/web/20251011105326/https://www.vice.com/en/article/an-ai-generated-artwork-won-first-place-at-a-state-fair-fine-arts-competition-and-artists-are-pissed/>
- 5 Jessica He, Stephanie Houde, and Justin D. Wiesz. *Which contributions deserve credit? Perceptions of attribution in human-AI co-creation*. *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–18, 2025. doi: <https://doi.org/10.1145/3706598.3713522>
- 6 Suzanne V. Wilson, Maria Strong, and Jordana Rubel. *Théâtre D’opéra Spatial Review Board Decision Letter*. United States Copyright Office Review Board, 2023. <https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf>

4.5 AI Literacy and Education in Design and Creative Practices

Samangi Wadinambiarachchi (The University of Melbourne, AU), Tilman Dingler (TU Delft, NL), Heekyoung Jung (University of Cincinnati, US), and Maria Luce Lupetti (Polytechnic University of Torino, IT)

License © Creative Commons BY 4.0 International license

© Samangi Wadinambiarachchi, Tilman Dingler, Heekyoung Jung, and Maria Luce Lupetti

Artificial intelligence is increasingly embedded in design and creative education and hence reshapes how students ideate, iterate, and produce work. While these tools offer clear benefits in terms of speed and experimentation, they also challenge long-standing pedagogical assumptions about the importance of process, authorship, and learning in studio-based contexts. Educators are therefore faced with a dual task: mitigating the risks AI poses to reflective, process-oriented design practice while simultaneously exploring how its capabilities can be meaningfully integrated into curricula. This report reflects on these tensions through the lens of design education, drawing on observations from a focused working group to identify emerging challenges, opportunities, and educational strategies for fostering critical, transparent, and sustainable uses of AI in creative classrooms.

4.5.1 AI and the Erosion of Process-Oriented Design Thinking

AI is rapidly diffusing into creative classrooms, creating a significant instability across both pedagogical and andragogical frameworks. While AI tools make production easier and lower the barriers to visualisation, creative processes tend to collapse by acting as a catalyst for premature idea synthesis [1]. This reinforces students' urge to jump immediately into solutions, bypassing the critical, process-focused approach essential to creative thinking. Therefore, the value of exercising critical judgment and iterating refinements, which is essential to the creative design process and practice, is on the verge of becoming diminished.

4.5.2 Ethical Uncertainty, Access Inequality, and Shifting Design Skillsets

Beyond these procedural shifts, as design educators, we have observed emerging behaviours, such as students struggling to distinguish when and how to disclose their approaches in accordance with the ethical appropriation of AI. Furthermore, access to AI tools has become a significant concern due to paywall restrictions and high subscription costs, making it hard to ensure equitable access to all students. There is also widespread confusion about which traditional design skills to continue practising, which new design skills to introduce, and how and when to integrate them with AI.

4.5.3 Making AI-Supported Design Processes Visible Through Prompt and Iteration Tracking

In group 5, we explored how design and creative education are affected by AI and considered measures to mitigate its negative effects while maximising AI's potential. We discussed ways to observe existing workflows and recognise interaction patterns to inform educational practices. In doing so, we identified that prompt histories and process logs can serve as reflective tools, revealing how the design process evolves by capturing struggles, inspirations, and breakthroughs. Similar to what classroom software like <https://www.khanacademy.org> allows in terms of classroom management and student monitoring, we envision a tool that keeps track of student's design iterations and the prompting that goes into each. This

makes students' AI-supported design processes visible, traceable, and discussable over time. Instead of focusing primarily on final outcomes, the dashboard documents how designs evolve alongside the prompts that informed them, treating prompting as a legitimate and assessable design practice. By linking iterations of a design to specific interactions with AI, the system supports reflection on how ideas emerge, shift, and mature, while enabling instructors to provide targeted feedback on students' prompting strategies and intentional use of AI. At the same time, it offers ways to address attribution and accountability, particularly in collaborative projects, by clarifying individual contributions and encouraging transparent disclosure of AI use.

4.5.4 Re-Centering the Creative Journey Through Process Documentation and AI Scaffolding

To restore a focus on the creative journey, we further discussed incorporating strategies such as reintroducing design journals, “blue books” to mandate process documentation. This emphasis on process documentation and reflection, both oral and written, can form a core component of evaluation methods and criteria of studio practice. Furthermore, we extended our discussion around scaffolding the use of appropriate AI tools in studio-based learning; for example, by understanding their constraints and capabilities; distinguishing AI as material, tool, and creative environment; and strategically integrating AI-enabled design processes across the curriculum to bridge foundational and emerging practices.

4.5.5 Toward Shared Educational Infrastructure and Sustainable AI Practices

We also discussed the need to identify the best approaches to develop automated workflows that support evaluation and EduTech development. Our discussions concluded with a proposal to create a platform with educational resources to raise awareness of best practices for AI use in educational contexts. This initiative could evolve into an open forum for educators to discuss emerging issues and engage diverse stakeholders, including educators, students and others, to outline AI usage policies in creative classrooms. Finally, this effort seeks to find a collective solution and to design a generative toolkit to foster sustainable appropriation of AI in creative and design education.

4.5.6 Strategic AI Containment in Design Education

Alongside efforts to integrate AI meaningfully into design education, we identified the need for deliberate containment strategies that protect core aspects of reflective and critical design practice. Especially in the context of studio-based learning, this includes intentionally reintroducing or strengthening assessment formats that resist full AI delegation, such as handwritten or in-class examinations, oral exams, live design critiques, and process-focused design journals. These approaches do not reject AI outright, but instead create pedagogical moments where students must articulate intent, judgment, and decision-making without relying on automated generation. By selectively constraining AI use, educators can counteract tendencies toward premature solution synthesis and ensure that foundational skills, such as critical reflection, reasoning, and verbalisation of design choices, remain central to learning outcomes. Framed in this way, AI containment can become an active instructional strategy that encourages slowing down the AI-fueled design process and adds reflective depth.

References

- 1 Habib, S., Vogel, T., Anli, X., & Thorne, E. *How does generative artificial intelligence impact student creativity?* *Journal of Creativity*, 34(1), 100072, Elsevier, 2024. doi: <https://doi.org/10.1016/j.yjoc.2023.100072>

5 Performance

5.1 The Agents

Dave Murray-Rust (TU Delft, NL), Joseph Lindley (Lancaster University, GB), Philippe Pasquier (Simon Fraser University – Surrey, CA), and Courtney Reed (Loughbourogh University London, GB)

License © Creative Commons BY 4.0 International license

© Dave Murray-Rust, Joseph Lindley, Philippe Pasquier, and Courtney Reed

URL <https://vimeo.com/1136322919/7df2f6e7a5?share=copy&fl=sv&fe=ci>

Creativity and improvisation go hand in hand. How better to think about creativity than through improvisational practice? This performance brought together four members of the Dagstuhl Seminar with their own individual improvising practices in an ad-hoc multimedia exploration of inter-agencies between humans and technologies. Dave Murray-Rust uses the laptop as an instrument, delegating micro-decisions to algorithmic processes while the musical shaping stays with fingertip controls. Philippe Pasquier marshals collections of musical agents to work with and against musical impulses, inhabiting a space between conducting and performing, and layering drones using the Kenaxis live software. Courtney Reed starts with the expressive range and power of operatic voice, then dives into a sea of electronic alterations and additions. Joe Lindley works visually, using realtime input to shape a stream of computational imagery through increasingly abstracted yet coherent dreamlike terrains. Performing for the first time, without rehearsal, they created the piece here as an example of lived practice within the discursive frame of the seminar.

Participants

- Jesse Josua Benjamin
TU Eindhoven, NL
- Federico Bomba
Freie Universität Bozen, IT
- Bapiste Caramiaux
Sorbonne University – Paris, FR
- Sarah Ciston
Center for Advanced Internet
Studies – Bochum, DE
- Laura Devendorf
University of Colorado –
Boulder, US
- Tilman Dingler
TU Delft, NL
- Sarah Fdili Alaoui
University of the Arts –
London, GB
- Elisa Giaccardi
Polytechnic University of
Milan, IT
- Lynda Hardman
CWI – Amsterdam, NL
- Mona Hedayati
University of Antwerp, BE &
Concordia University –
Montreal – CA
- Janet Johnson
University of Michigan –
Ann Arbor, US
- Heekyoung Jung
University of Cincinnati, US
- Lone Koefoed Hansen
Aarhus University, DK
- Joseph Lindley
Lancaster University, GB
- Maria Luce Lupetti
Polytechnic University of
Torino, IT
- Christoph Maerz
DFKI –
Kaiserslautern, DE
- Dave Murray-Rust
TU Delft, NL
- Philippe Pasquier
Simon Fraser University –
Surrey, CA
- Courtney Reed
Loughbourogh University
London, GB
- Steven Rick
NVIDIA – Santa Clara, US &
MIT – Cambridge, US
- Irina Shklovski
University of Copenhagen, DK
- Vasiliki Tsaknaki
IT University of
Copenhagen, DK
- Libuše Hannah Veprek
Universität Tübingen, DE
- Samangi Wadinambiarachchi
The University of Melbourne, AU
- Justin D. Weisz
IBM TJ Watson Research Center
– Yorktown Heights, US
- Boyu Xu
Utrecht University, NL
- Hiromu Yakura
MPI für Bildungsforschung –
Berlin, DE

